First publ. in: IEEE Transactions on Systems, Man and Cybernetics : B 30 (2000), 6, pp. 821 - 834

Input Features' Impact on Fuzzy Decision Processes

Rosaria Silipo, *Member, IEEE,* and Michael R. Berthold, *Member, IEEE*

***Abstract*—Many real-world applications 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 fuzzy models can progressively disappear. An important part of the model interpretation lies on the evaluation of the effectiveness of the input features on the decision process. In this paper, we present a method that quantifies the discriminative power of the input features in a fuzzy model. The separability among all the rules of the fuzzy model produces a measure of the information available in the system. Such measure of information is calculated to characterize the system before and after each input feature is used for classification. The resulting information gain quantifies the discriminative power of that input feature. The comparison among the information gains of the different input features can yield better insights into the selected fuzzy classification strategy, even for very high-dimensional cases, and can lead to a possible reduction of the input space dimension. Several artificial and real-world data analysis scenarios are reported as examples in order to illustrate the characteristics and potentialities of the proposed method.**

***Index Terms*—Discriminitive power, feature importance, fuzzy models, information gain.**

I. Introduction

A. Interpretability of Decision Process

IN THE last several years, it has become increasingly common to collect and store large amounts of data from different sources, as described in [1, Ch. 1]. As a consequence, databases with higher dimension and bigger size have been obtained. In this paper, we deal with two typical research areas, where big high-dimensional databases have been developed: the analysis of medical signals and the automatic speech recognition problem.

The recording of electrocardiographic (ECG) signals, for example, moved to 24-hr and 12-lead just a few years ago. At the same time, the number of features extracted from each ECG record increased as well [2]. These days, the current tendency in medical databases is to collect heterogeneous data from many physiological sources and for long time periods. A very typical example for this new kind of data is the Apnea-ECG Database, which is downloadable from the PhysioNet web site (http://www.physionet.org/). This database consists of 70 records. Each record is typically 8 hr long and contains simultaneously recorded ECG and respiration signals.

R. Silipo is with the International Computer Science Institute (ICSI), Berkeley, CA 94704-1198 USA (e-mail: rosaria@icsi.berkeley.edu).

M. R. Berthold is with the Berkeley Initiative in Soft Computing (BISC), University of California, Berkeley, CA 94720 USA (e-mail: berthold@cs.berkeley.edu).

Other examples of big size databases are also available in the automatic speech recognition research field. The OGI Corpus [3] used in this study, for example, consists of responses to prompts spoken over commercial telephone lines by speakers of English, Farsi (Persian), French, German, Hindi, Japanese, Korean, Mandarin Chinese, Spanish, Tamil, and Vietnamese. It contains a total of 1927 calls: an average of 175 calls per language. Current systems for automatic speech recognition derive between 150–250 input features from the original signal and systems are being developed with an even higher number of input features [4].

Despite the efforts in this direction, the collection of more signals from different sources or the extraction of more input features does not always grant a better performance of further analysis procedures. If the newly introduced variables do not carry additional information, the system's performance cannot improve. Moreover, the analysis procedure itself becomes more complicated, and insights about the system's underlying structure become more difficult to achieve.

The interpretability of the decision process represents a key topic in modern data analysis scenarios and corresponds to the transparency of the model built to implement a given task. For example, if, in a given context, a data analysis technique does not show as good performance as other methods but offers a more informative representation of the underlying phenomenon and/or a clearer interpretation of the decision process, such a technique may represent a better decision support tool for the user than a technique that offers numerically superior performance but is harder to interpret.

An important part of the interpretation of a decision process lies in the assessment of the influence of its input features on the final decision, that is, on the assessment of how much the implemented model relies on a given input feature to perform the desired task.

Much work has been done in the area of discovering feature importance, mainly under the umbrella of feature selection. The most commonly used methods stem from the area of probabilistic decision trees, particularly ID3 [5] and its continuous extension C4.5 [6]. Following the theory of entropy maximization in probabilistic decision trees, some merit measures have been defined on the basis of a statistical model of the system [1, Ch. 3], [7]. The estimation of the involved probabilities, however, requires a precise definition of the input parameters and a clear identification of the output classes. In many real-world applications, an inaccurate description of the input parameters and doubtful members of the output classes unavoidably alter the event frequencies used for probability estimation. In addition, the estimation of a probabilistic model may be computationally expensive for high-dimensional input spaces.

Recently, many data analysis techniques make use of the easy interpretability and low computational expenses of fuzzy logic, such as fuzzy rules induction, fuzzy decision trees, etc. [1, Ch.

8]. The concept of fuzzy sets was introduced in [8] with the purpose of a more efficient, although less detailed, description of real-world events, by allowing an appropriate amount of uncertainty into the data description. A number of simple and computationally inexpensive methods are now available to automatically construct a model from a set of training examples [9]–[11]. For example, a fuzzy extension to ID3, which requires predefined granulation on all input features, was proposed in [12].

If the particular problem does not require very complex decision borders among the output classes, fuzzy models produce a reasonable amount of fuzzy rules that offer sufficiently reliable performances and, for a low-dimensional input space, are relatively easy to interpret. Many real-world applications present very high-dimensional input spaces and require very complex decision borders. Because of that, the number of fuzzy rules can proliferate, and the easy interpretability of fuzzy models can progressively disappear. In this case, the introduction of an automatic description of some of the characteristics of the fuzzy model would improve its interpretability to the user.

B. Input Features' Impact on Fuzzy Models

One important characteristic that describes the implemented fuzzy model consists of the impact of the input features on the final decision process. The goal of this work is to define a strategy to automatically quantify the influence of the input features on the fuzzy model. Such influence could be measured by estimating and comparing the information contained in the fuzzy model before and after using that input feature for the analysis. In information theory, the information associated with a given event is measured by means of its entropy. Dealing with fuzzy models, the concept of fuzzy entropy [9]–[14] could be used for the same purpose.

Based on fuzzy set theory, fuzzy entropy has been defined to measure the degree of fuzziness/uncertainty of the model in fitting the desired input/output mapping with respect to the training examples [9]–[11], [13], [14]. Such a measure of information can be computationally expensive and time consuming if very large data sets are used. Moreover, it would characterize the input features in terms of the faithfulness of the model to the training examples and would fail to give a description of their discriminative power in separating the output classes.

The method proposed in this paper investigates only the fuzzy model, which is, in general, a mere summary of the training examples. Indeed, if the training set contains a sufficient number of examples—that is, if the resulting fuzzy model is sufficiently general and accurate—an analysis *a posteriori* of the fuzzy model's characteristics will reflect information about the input space. In addition, by concentrating only on the fuzzy model, the corresponding analysis will be computationally easier and faster.

Thus, the information available in the fuzzy model is derived solely on the basis of its fuzzy rules. The original fuzzy model is split into a certain number of fuzzy submodels, according to the linguistic values of the given input feature, and the average information contained in these fuzzy submodels is compared with the information in the original fuzzy model. The resulting information gain quantifies the information extracted from the model after using this input feature for the analysis and characterizes its discriminative power inside the model. The input dimension with highest information gain defines the most discriminative input feature, according to the analyzed fuzzy model. Unlike the greedy behavior of the probabilistic decision tree algorithms, this method investigates the cuts on each input feature—not one after the other but all together in parallel—which enables it to find also nonbinary splits.

Theoretically, both positive and negative information gains are possible. In the first case, the input feature has a positive impact on the decision process. In the second case, the input feature worsens the system's performance. In practice, if a sufficient amount of data is available, the classification method should learn to neglect the unreliable input features. Thus, only positive or zero information gains can be obtained. Because of machine precision errors and of the imperfections of the learning procedure, some negative close-to-zero information gains might arise. Because of their low absolute values, in the following analysis, we will ignore negative information gains and report them to zero.

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 fuzzy classification strategy, especially for very high-dimensional input spaces and, consequently, a possible reduction of the input dimension.

The structure of the paper is the following. After describing the need of interpretable decision processes in Section I-A, we illustrate the goal of the paper and the general idea of the method in Section I-B. In Section II, we define how to measure the information contained in the fuzzy model. Then, in Section III, we use this measure of information to characterize the system before and after a given input feature is used for classification. The resulting information gain is described in Section III-B. In Section IV, some artificial data are analyzed to show the potentiality of the proposed method. Finally, in Section V, three real-world applications are investigated. The first one (Section V-A) uses the IRIS database, which represents a common platform for the evaluation of machine learning algorithms. The second application deals with the automatic detection of prosodic stress in spoken American English (Section V-B) and tries to rank the most commonly used input features in terms of impact on the decision process. The last real-world application (Section V-C) investigates whether removing ECG measures with low information gains improves the performance of a fuzzy system trained to discriminate among different kinds of arrhythmic beats. Section VI concludes the paper.

II. Fuzzy Information Measures

Fuzzy models represent a particular version of rule sets, where some uncertainty or *fuzziness* is allowed, so that a given input pattern $\vec{x}$, which is composed of features $x_1, \cdots, x_j, \cdots, x_n$, belongs up to a given degree (*membership degree*) to a certain output class C_i ($1 \leq i \leq m$) [8]. Thus, the set of rules implementing this kind of input/output mapping consists of a set of membership functions $\mu_{C_i}(\vec{x}) \in [0, 1]$ that associate input pattern $\vec{x}$ to output class C_i by means of membership degree $\mu_{C_i}(\vec{x})$.

Given a number m of output classes C_i and an n-dimensional input space, numerous algorithms exist, which derive a set of

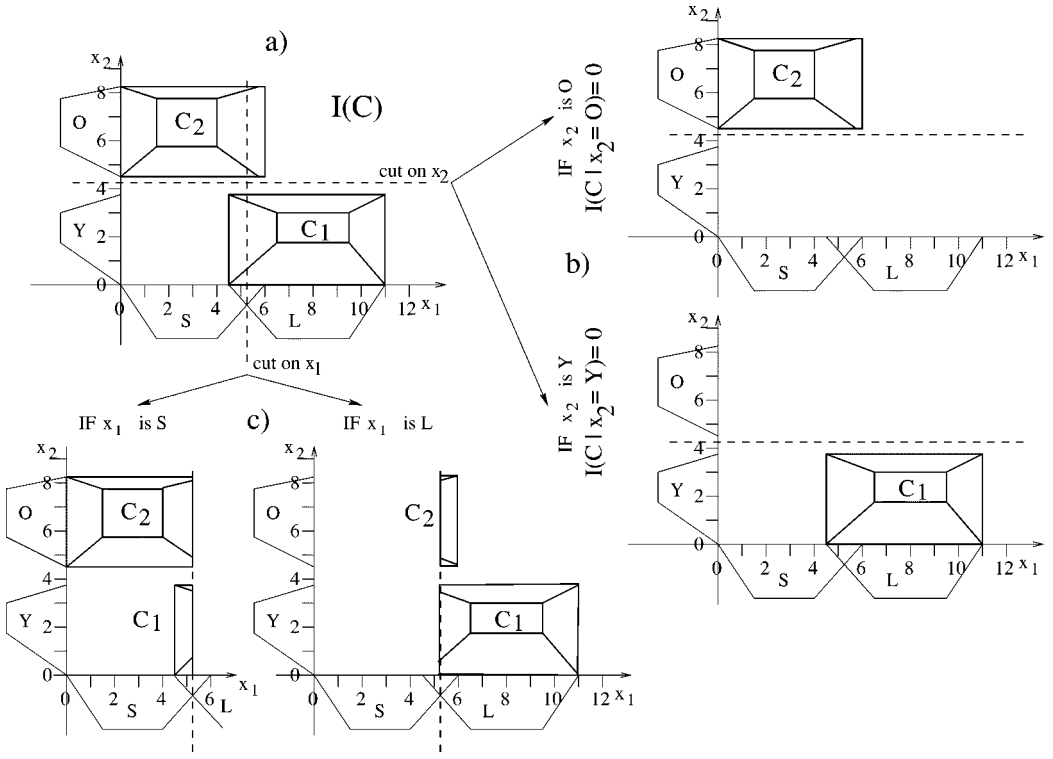

Fig. 1. (a) Example of a two-class fuzzy model on a 2-D input space. (b) and (c) Submodels generated by cutting the original fuzzy model in (a) along input feature (b) x_2 and (c) x_1.

N_R fuzzy rules $\{R_k\}$, $k = 1, \cdots, N_R$, mapping the n-dimensional input into the m-dimensional output space. In particular, we used the fuzzy clustering algorithm proposed in [18]. This algorithm adapts existing fuzzy rules to new input patterns and introduces new rules when necessary. The algorithm is guaranteed to converge, and an upper bound on the number of the generated rules can also be introduced [18]. In Fig. 1(a), an example is reported with a 2-D input space $\{x_1, x_2\}$, two output classes C_1, and C_2 and with trapezoids as membership functions.

A. Average Membership Degree

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 (1) represents the *average degree of membership* of input patterns $\vec{x}$ to output class C_i over the whole domain $D \subset \mathbb{R}^n$.

$$V(C_i) = \frac{\int_{D \subset \mathbb{R}^n} \mu_{C_i}(\vec{x})\, d\vec{x}}{\int_{D \subset \mathbb{R}^n} d\vec{x}}. \tag{1}$$

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 that takes value $+1$ everywhere on the input domain 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 D. This average membership degree $V(C_i)$ [see (1)] represents a first rough description of the impact of membership function $\mu_{C_i}(\vec{x})$ on the final decision process without taking into account the training examples from which $\mu_{C_i}(\vec{x})$ originates.

In order to quantify the information contained in the whole set of fuzzy rules $\{R_k\}$, all average membership degrees from the different membership functions should be considered at the same time. The goal of this section is to associate different configurations of average membership degrees to fuzzy models with different informative contents. In particular, some mathematical operator could be applied to $V(C_i)$ to distinguish between fuzzy models with only membership functions of one class (no information) from fuzzy models with membership functions of a high number of output classes (high information).

In information theory, a number of functions, such as the entropy or the Gini function, have been established in the past to play this role in a probabilistic context [7], [15]. However, they cannot be applied straightforward to the average membership degree $V(C_i)$ because of the requirement that the object variable sums up to $+1$ across the m output classes C_i. Unlike for probability, this is not necessarily true for the average membership degrees $V(C_i)$, due to the nonnormalized nature of fuzzy sets.

A solution to this problem consists of using the *relative average membership degree* [see (2)] $v(C_i)$ to output class C_i instead of the average membership degree $V(C_i)$ [16]:

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

The variable $v(C_i)$, with $i = 1, \cdots, m$, now sums up to $+1$ across the output classes, and the traditional information functions can be applied.

In general, a number $Q_i > 1$ of membership functions is necessary to represent each output class C_i. Each one of these membership functions is related to an output region C_i^q and, therefore, will be indicated as $\mu_{C_i}^q(\vec{x})$. Thus, the average membership degree to class C_i corresponds to the average membership degree to the union of the corresponding output regions C_i^q.

The average membership degree to the union and the intersection of fuzzy sets derives straightforward from the usual min/max-definitions of intersection and union of fuzzy sets [17]. In particular, the average membership degree to the union of two fuzzy sets C_i^r and C_i^s can be derived as the sum of the average membership degrees to the two fuzzy sets alone, taking into account their intersection only once [see (3) and (4)].

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

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

If the two membership functions $\mu_{C_i}^r(\vec{x})$ and $\mu_{C_i}^s(\vec{x})$ do not overlap, that is, $\forall \vec{x}$: $\min_{r,\,s}\{\mu_{C_i}^r(\vec{x}),\, \mu_{C_i}^s(\vec{x})\} = 0$, the expressions in (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 to their union as the sum of their average membership degrees and taking care of including their intersection only once.

$$\begin{aligned} V(C_i) &= V\left(\bigcup_{q=1}^{Q_i} C_i^q\right) \\ &= \frac{\displaystyle\int_{D\subset\mathbb{R}^n} \max_q\left\{\mu_{C_i}^q(\vec{x})\right\}\, d\vec{x}}{\displaystyle\int_{D\subset\mathbb{R}^n} d\vec{x}} \\ &= \sum_{q=1}^{Q_i}\left[V(C_i^q) - \sum_{h=q+1}^{Q_i} V\left(C_i^q \cap C_i^h\right)\right]. \end{aligned} \tag{7}$$

If the usual trapezoids are adopted as membership functions, the average membership degree to each fuzzy subset C_i^q becomes particularly simple to calculate [17], as shown in (8), where h^q is the trapezoid height, and $\langle \vec{a}_i^q,\, \vec{b}_i^q,\, \vec{c}_i^q,\, \vec{d}_i^q \rangle$ are the coordinate vectors of its vertices in the n-dimensional input space.

$$\begin{aligned} V(C_i^q) &= V(\langle \vec{a}_i^q,\, \vec{b}_i^q,\, \vec{c}_i^q,\, \vec{d}_i^q \rangle) \\ &= \frac{\frac{1}{2}\left(\displaystyle\prod_{j=1}^{n}(d_{ji}^q - a_{ji}^q) + \prod_{j=1}^{n}(c_{ji}^q - b_{ji}^q)\right) h^q}{\displaystyle\int_{D\subset\mathbb{R}^n} d\vec{x}}. \end{aligned} \tag{8}$$

B. Fuzzy Information Measures

As we have already described in the previous subsection, we could take the quantity $v(C_i)$ as the basic unit to quantify the information available in a fuzzy 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 (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 $v(C_i)$ 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 [1, Ch. 3], that is

1) at its maximum if all the output classes are equally possible in average on the input space $D \subset \mathbb{R}^n$, i.e., $v(C_i) = (1/m)$ for $i = 1, \cdots, m$, m being the number of output classes;
2) at its minimum if only one output class C_i exists, i. e. in case $v(C_i) \neq 0$ and $v(C_j) = 0$ 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 favorite class is.

In order to produce a measure of the global information $I(C)$ of the fuzzy model with output space $C = \{C_1, \cdots, C_m\}$, the traditional functions employed in information theory—as the entropy function $I_H(C)$ [see (9)] and the Gini function $I_G(C)$ [see (10)] [1, Ch. 3], [7]—can be applied to the relative average membership degrees $v(C_i)$ of the output classes as

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

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

and the following conditions 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 ($v(C_i) \gg v(C_j)$, $j \neq i$) produces the same amount of information, independent of which one is the favorite class. That is, the defined information functions of variable $v(C_i)$ [see (9) and (10)] are symmetric.

In both cases, the entropy and Gini functions $I(C)$ represent the information intrinsically available in the fuzzy model. The classification process aims to extract such information for the user's needs. Not all the input features, however, are effective the same way in extracting and representing this information. The goal of this paper is to make explicit the dimension of the input space that is the most effective in recovering the intrinsic information $I(C)$ contained in the fuzzy model.

III. 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. In particular, such information gain can be expressed as the relative difference between the intrinsic information available in the system before—$I(C)$—and after—$I(C|x_j)$—*using* input feature x_j for the fuzzy analysis [7]. In the following, we define what the use of x_j corresponds to and how to measure the information left in the system after input feature x_j has been exploited for the analysis (Sections III-A and III-B, respectively).

A. Key Points on Input Dimension x_j

Let us suppose that input space $D \subset \mathbb{R}^n$ is related to the output classes by means of a number N_R of membership functions $\mu^q_{C_i}(\vec{x})$ with $q = 1, \cdots, Q_i$ membership functions for each output class C_i, $i = 1, \cdots, m$ output classes, and $N_R = \sum_{i=1}^{m} Q_i$ fuzzy rules.

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. From a risk minimization point of view, the optimal classification thresholds on a given input dimension x_j are located at the intersection points of contiguous membership functions of different output classes.

Let us restrict our analysis to a 1-D problem. In Fig. 2, an example with two output classes on a 1-D space x is reported. Let us choose a discrimination threshold x^* to separate class C_1 from class 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

$$F = \int_{\hat{C}_2} \mu_{C_1}(x)\,dx + \int_{\hat{C}_1} \mu_{C_2}(x)\,dx. \qquad (11)$$

The optimal classification threshold x^* refers to the minimum degree of falseness (F) of the whole classification process, that is, to the minimum intersection volumes $\mu_{C_1} \cap \hat{C}_2$ and $\mu_{C_2} \cap \hat{C}_1$. After minimizing (11), the optimal threshold x^* is found at the intersection point of the two membership functions x^*: $\mu_{C_1}(x^*) = \mu_{C_2}(x^*)$.

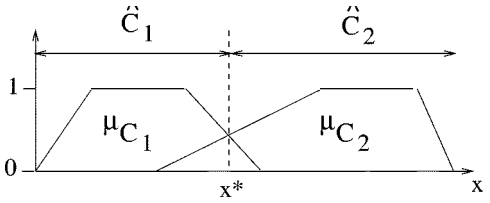

Fig. 2. Fuzzy representation of a 1-D input space with two output classes.

If trapezoids are adopted as membership functions of the fuzzy model, the optimal threshold between two contiguous trapezoids of different output classes is assumed to be located

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 ($\mu \equiv 1$, which are 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.

The definition of a set of thresholds based on the risk minimization approach is typical of the statistical classification strategies. In a fuzzy context, input x may belong to both output classes C_1 and C_2. To be fully in line with the fuzzy classification strategy, a different threshold system should be developed that takes into account the attribution of pattern x to multiple classes. However, such a system would be more complex and computationally expensive than the one based on the risk minimization approach. In addition, in this paper, we identify the effectiveness of a given input feature with the separability of the output classes along its dimension, which is well represented by the risk minimization-based threshold system. Thus, we retained the set of thresholds defined in this section for the quantification of the input features' impact because it is a sufficiently accurate and leads to an algorithm with lower computational load. These thresholds are used only to quantify the separability of the output classes along a given input dimension in the definition of the fuzzy feature merit measure. We adopted the traditional fuzzy classification strategy that allows each input pattern to belong to more output classes at the same time to test the fuzzy models.

B. Information Gain

The discrimination of the output classes along input feature x_j leads to the definition of a set of optimal cuts that separate the $F_j \leq N_R$ contiguous trapezoids on this input dimension, as discussed in the section above. After introducing the upper and lower boundary of x_j's range in this set of optimal cuts, a number of linguistic values L_k ($k = 1, \cdots, F_j$) can be defined for input feature x_j as the intervals between two consecutive cuts.

Let us concentrate on one x_j's linguistic value L_k per time. To consider $x_j = L_k$ corresponds to the isolation of one stripe of the input space where x_j falls into linguistic value L_k. In this stripe $x_j = L_k$, new membership functions $\mu^q_{C_i}(x_j = L_k)$ are derived as intersections of the original membership functions $\mu^q_{C_i}(\vec{x})$ with the stripe derived from $x_j = L_k$. Based on these new membership functions $\mu^q_{C_i}(x_j = L_k)$, the information contained in this stripe can be measured $I(C|x_j = L_k)$, as

expressed in (12) and (13) according to the information functions in (9) and (10), respectively.

$$I_H(C|x_j = L_k) = - \sum_{i=1}^{m} v(C_i|x_j = L_k) \log_2(v(C_i|x_j = L_k)) \quad (12)$$

$$I_G(C|x_j = L_k) = 1 - \sum_{i=1}^{m} (v(C_i|x_j = L_k))^2 \quad (13)$$

with

$$v(C_i|x_j = L_k) = \frac{V(C_i|x_j = L_k)}{\sum_{h=1}^{m} V(C_h|x_j = L_k)}. \quad (14)$$

$I(C|x_j = L_k)$ measures the information still available in the stripe extracted from the original fuzzy model under the condition that x_j falls inside linguistic value L_k. The average measure of the information contained in all stripes $x_j = L_k$ for $k = 1, \cdots, F_j$ [see (15)] represents the measure of the information still available in average in the fuzzy model after input feature x_j has been exploited for the fuzzy analysis $I(C|x_j)$.

$$I(C|x_j) = \frac{1}{F_j} \sum_{k=1}^{F_j} I(C|x_j = L_k). \quad (15)$$

The relative difference, as expressed in (16), between the measure of the information originally available in average in the fuzzy model $I(C)$ and the measure of the information still available after the use of input feature x_j, $I(C|x_j)$ produces the corresponding information gain.

$$g(C|x_j) = \frac{I(C) - I(C|x_j)}{I(C)}. \quad (16)$$

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 $g(C|x_j)$ [see (16)]. The input features producing the highest information gains are the most effective in the adopted model to separate the training data 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. The proposed information gain can then be adopted as a fuzzy feature merit measure.

C. Example

In Fig. 1, an example is shown with a 2-D input space, two output classes, and trapezoids as membership functions. In Table I, the absolute and relative average membership degrees of the two classes are reported, and based on these values, the information $I_H(C)$ and $I_G(C)$ that is intrinsically available in the model is measured by means of (9) and (10).

The discrimination of the two output classes can now be performed along input dimension x_1 or along input dimension x_2. From Fig. 1, we can easily see that a cut between the two membership functions on dimension x_2 [see Fig. 1(b)] produces a better separation than a cut on dimension x_1 [see Fig. 1(c)]. That is, the analysis on dimension x_2 should offer a higher gain in information than the analysis on dimension x_1.

TABLE I
AVERAGE MEMBERSHIP DEGREES AND THE INFORMATION MEASURES FOR THE 2-D EXAMPLE IN FIG. 1

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 II
$I(C|x_j)$ AND $g(C|x_j)$ FOR THE EXAMPLE IN FIG. 1

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

To verify this hypothesis, the average information still available in the system $I(C|x_1)$ and $I(C|x_2)$ is measured, respectively, after dimension x_1 and x_2 have been used for the classification. These information measures are reported in Table II together with the corresponding 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, $g(C|x_1) < g(C|x_2)$ (see Table II), 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 out on input feature x_1. We could reach the same conclusion using $I(C|x_1) > I(C|x_2)$. However, a measure of merit based on the gain function produces clearer results than the direct use of the information parameter $I(C|x_j)$.

IV. ARTIFICIAL DATA EXAMPLES

In this section, we analyze some artificial examples to show how the information gain defined in Section III-B characterizes the effectiveness of the input features for the required fuzzy input/output mapping. For all the examples reported in this study, we used the fuzzy clustering algorithm proposed in [18] to build the set of fuzzy rules approximating the classification task at hand and trapezoids as membership functions.

A. Fixed Output Classes

The first example refers to a three-dimensional (3-D) problem with four output classes. The projection of such input space on a 2-D plane is shown in Fig. 3. Random values are generated for the third dimension of all patterns. Fig. 3 shows that a correct classification of all input data cannot be obtained on the basis of only one input feature. Both input features x_1 and x_2 seem

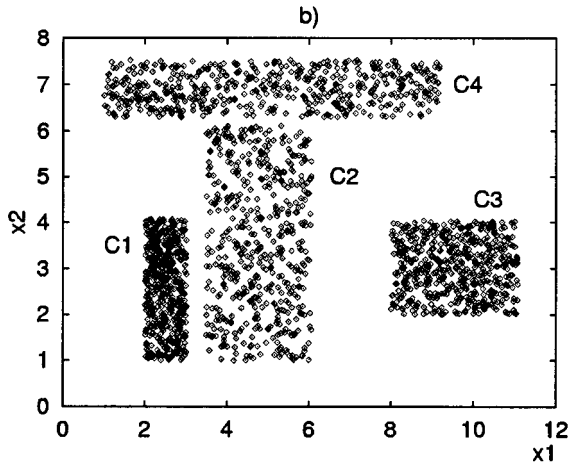

Fig. 3. Two-dimensional projection of a 3-D input space with four output classes. The third dimension consists of random values for all output classes.

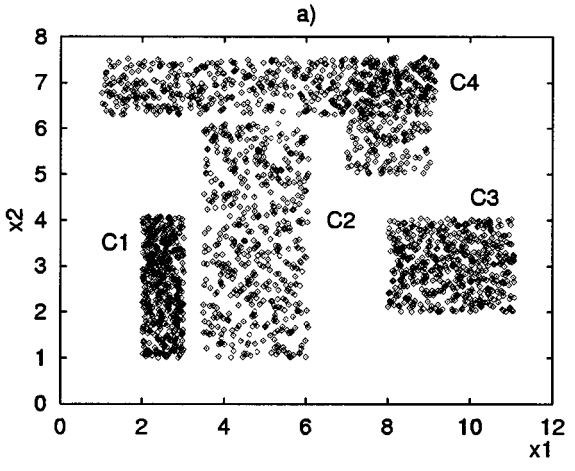

Fig. 4. Variation of the input space with four output classes in Fig. 3. The input space is 3-D, and its third dimension x_3 consists of random values for all output classes.

TABLE III
INFORMATION MEASURES—$I_H(C)$, $I_G(C)$, $I_H(C|x_y)$, AND $I_G(C|x_y)$—AND INFORMATION GAINS—$g_H(C|x_y)$ AND $g_G(C|x_y)$—FROM THE FUZZY MODEL CONSTRUCTED ON THE INPUT SPACE DESCRIBED IN FIG. 3

$I_H(C)$	1.87		
$I_G(C)$	0.71		
dimension	x_1	x_2	x_3
$I_H(C\|x_y)$	0.90	1.20	1.87
$I_G(C\|x_y)$	0.41	0.51	0.71
$g_H(C\|x_y)$	0.52	0.36	0.00
$g_G(C\|x_y)$	0.42	0.28	0.00

TABLE IV
INFORMATION MEASURES—$I_H(C)$, $I_G(C)$, $I_H(C|x_y)$, AND $I_G(C|x_y)$—AND INFORMATION GAINS—$g_H(C|x_y)$ AND $g_G(C|x_y)$—FROM THE FUZZY MODEL CONSTRUCTED ON THE INPUT SPACE DESCRIBED IN FIG. 4

$I_H(C)$	1.85		
$I_G(C)$	0.70		
dimension	x_1	x_2	x_3
$H(C\|x_y)$	0.90	1.30	1.85
$G(C\|x_y)$	0.42	0.55	0.70
$g_H(C\|x_y)$	0.51	0.30	0.00
$g_G(C\|x_y)$	0.40	0.22	0.00

to be necessary for this purpose. A fuzzy model is implemented using these data points as training set. The corresponding information measures and gains are reported in Table III for every input dimension.

Let us concentrate on the information gain values in Table III. The third dimension (x_3) contributes to the overall classification task with an information gain equal to 0.0, as was to be expected, because of its random values in all four output classes. However, 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 any input dimension alone. Input features x_1 and x_2 present similar 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, whereas three output classes can be separated along x_1. Thus, 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 order to test the strength of the fuzzy information gain parameter in quantifying the discriminative power of the input features, the input space depicted in Fig. 3 was slightly changed in Fig. 4 so that class C_4 overlaps with class C_2, even on the x_2 axis. Therefore, the discriminability of the output classes should decrease mainly on x_2 and slightly on x_1 with respect to the example in Fig. 3.

The new information gain values are reported in Table IV. 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 in the previous example.

B. Moving Output Classes

In the previous subsection, we have shown that the proposed information gain is able to quantify the discriminative power of the input features in fuzzy models 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.

Let us start with a configuration in a 2-D input space, where two output classes are completely separable along one input dimension and completely overlapping along the other, as described in Fig. 5. In the next snapshots, one of the two output classes (C_1) is progressively shifted 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 information gains referring to the initial configuration of the input space (see Fig. 5) are reported in the first row of Table V. As was to be expected, an information gain close to

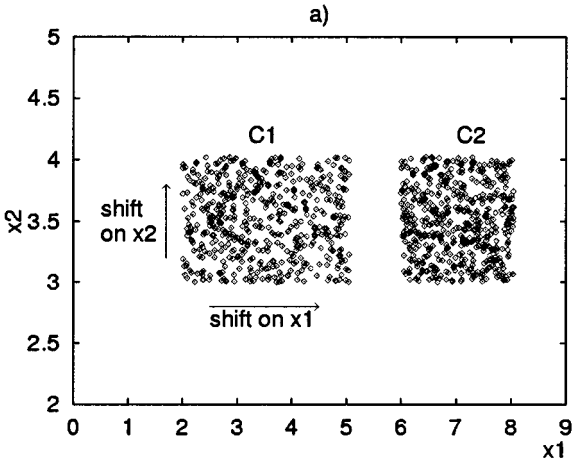

Fig. 5. Two-dimensional input space with two output classes progressively overlapping on one input dimension.

1.0 describes an almost perfect separability of the two output classes along x_1, whereas 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 toward class C_2 along the x_1-axis with a step Δx_1, whereas their x_2 coordinate stays constant. The corresponding information gains are reported in the following rows in the upper part of Table V.

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. The two output classes begin to overlap for $\Delta x_1 = +1.5$, and after that, 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 as well. Continuing to shift C_1 class' 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, as does the information gain, until values close to 1.0 are re-established, 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 increasingly approach the unitary value. However, at $\Delta x_1 = +6.5$, a small decrease in the information gain is observed, even though the two output classes are more separated than for $\Delta x_1 = +6.0$. In this case, the adopted fuzzy learning algorithm builds less steep trapezoids than for closer output classes because of the nonexistence of conflict points [18]. 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 also does not change from the first row of Table V.

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

TABLE V
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 FIG. 5 AND SHIFTING CLASS C_1 TOWARDS CLASS C_2 ALONG x_1 WITH 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

Progressively increasing 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 V. 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 gains ($g_H(C|x_2) = 0.0$ and $g_G(C|x_2) = 0.0$) are 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 is reached 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 input dimension x_1 and x_2.

These results show that the proposed fuzzy feature merit measure is able to detect the dimension with maximum information content for different configuration of the output classes in the training set. An information gain close to 1.0 is shown on those

input dimensions where an almost 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. The fuzziness of the system does not allow an information gain of 1.0 when the two output classes are not overlapping anymore but are still 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 of 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 [18].

V. 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 different input dimensions for artificially created data. In this section, real-world data are investigated.

A. IRIS Database

The first experiment is performed on the IRIS database. This database is relatively small, and the results cannot be easily generalized. On the other hand, it is a commonly used database, which enables a comparison with other similar techniques.

The IRIS database contains data for three classes of iris plants (iris setosa, iris virginica, and iris versicolor). The first class is linearly separable, whereas the last two classes are not. 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 [19], where a detailed description of the 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 versicolor).

The fuzzy clustering algorithm [18] is trained by using the whole database as training set. The corresponding information gains for each input feature are calculated and reported in Table VI. The third and the fourth input parameter (x_3 and x_4) exhibit very high information gains, whereas 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 is in agreement with that which is described in [19]. In [20], a statistical correlation measure of the output classes with the input features is also reported. Parameters x_3 and x_4 have a very high correlation with the output classes, whereas x_1 and x_2 are associated with a much lower correlation value. This confirms the hypothesis of a more informative character of x_3 and x_4 derived from the fuzzy feature merit measures in Table VI.

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.

TABLE VI
Information Gains $g_H(C|x_y)$ and $g_G(C|x_y)$ of the Input Features in the Iris Database. x_1 = Sepal Length; x_2 = Sepal Width; x_3 = Petal Length; x_4 = Petal Width

$I(C)$		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

B. Stress Detection in Spoken American English

Prosodic stress is an integral component of spoken language, particularly for languages such as English that so heavily depend on this parameter for lexical, syntactic, and semantic disambiguation. Even though it is by now quite generally accepted [21]–[23] that prosodic stress depends mainly on amplitude, duration, and pitch of the vocalic nuclei of syllables in spoken American English, the role played by each one of these basic parameters is still controversial.

In this section, the fuzzy information gain described in Section III-B is applied to the problem of automatic detection of prosodic stress in spoken American English to ascertain the role pertaining to each one of these basic parameters for reliable stress recognition.

The basic parameters, characterizing each vocalic nucleus, are quantified as follows.

- **Duration:** Inside a speech file, the *duration* of the kth 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 kth vocalic nucleus.
- **Average Pitch:** The *average pitch* P_k refers to the average value of the fundamental frequency $f_0(t)$ inside the kth vocalic nucleus. Fundamental frequencies $f_0(t)$ are estimated on the basis of the autocorrelation function of quarter of octave spectral channels, as described in [24].
- **Pitch Range:** The *pitch range* P_k refers to the range of values of the fundamental frequency $f_0(t)$ inside the kth vocalic nucleus.

Diphthongs, such as "ay," "oy," and "er" present a longer duration than plain vowels and, because of that, are divided in two parts. For the same reason, artificially elongated vowels that are longer than 250 ms and 400 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, average pitch, and pitch range. In order to normalize this variance among speakers, those features are expressed in terms of variance units from the mean value of their probabilistic distributions inside each utterance [24].

To provide a reference platform for the system's performance, two trained linguists separately hand labeled two

TABLE VII
NUMBER OF FILES FROM THE OGI STORIES DATABASE LABELED BY EACH TRANSCRIBER

voices	first transcriber	second transcriber	both
men	49	38	5
women	34	13	5
total	83	51	10

TABLE VIII
IN THE FIRST THREE COLUMNS: AGREEMENT OF TRANSCRIBER # 1 VERSUS. TRANSCRIBER # 2. IN THE LAST THREE COLUMNS: AGREEMENT OF TRANSCRIBER # 2 VERSUS. TRANSCRIBER # 1. THE AGREEMENT PERCENTAGES ARE CALCULATED ON ALL COMMON FILES. $S+$ PRIMARY, $S-$ MINOR STRESSED, N UNSTRESSED VOWELS

Transcr. # 1 vs. # 2			Transcr. # 2 vs. # 1		
% agreement			% agreement		
S+	S-	N	S+	S-	N
90	67	84	78	57	93

different subsets of the American English component of the OGI Stories Corpus [3] in terms of prosodic stress (see Table VII). The OGI corpus contains 50–60-s files of spontaneous speech about any subject. Ten files—five men's and five women's voices—are common to both subsets. The stress annotations refer to primary stressed $(S+)$, other minor stressed $(S-)$, and unstressed syllables (N).

The agreement between the two transcribers on the common files is shown in Table VIII and will be used as a baseline for the system's performance. The first three columns of Table VIII refer to the agreement percentage of transcriber# 1 versus transcriber# 2 and the second three columns to that of transcriber# 2 versus transcriber# 1. Since only a two-level stress automatic classification (stressed versus unstressed syllables) is implemented, the agreement percentages in Table VIII are calculated accordingly. A stressed syllable labeled as $S+$ (or $S-$) by one transcriber is considered in agreement if the other transcriber labeled it also as either $S+$ or $S-$. The two transcribers roughly agree in recognizing primary stress $(S+: 90\text{–}78\%)$ versus unstressed syllables N: $84\text{–}93\%)$. Much more disagreement exists in recognizing minor stresses $(S-: 67\text{–}57\%)$.

From each subset of annotated files from the OGI database, two thirds of the files are used as a training set to implement a fuzzy model [18] that discriminates stressed ($S+$ and $S-$) versus unstressed (N) vocalic nuclei. The resulting fuzzy model is tested on the remaining one third of files and analyzed in terms of the discriminative power granted to each input feature (see Tables IX and X).

During the test phase, each membership function is weighted with the number of training patterns covered at the end of the training procedure. This helps to solve conflicts among membership functions, favoring the one representing the highest number of training patterns. 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 nonzero membership degrees. The percentage of correctly classified test patterns for each output class is defined as the sum of correct answers with respect to the number of test patterns of this output class.

The training and testing procedure is repeated using the Jackknife method. Two thirds of the files that are used as a training set, and the one third used as a test set, are cyclically exchanged in such a way as to obtain three different pairs of training and test sets. The average system's performance and input features information gains are calculated across the three pairs training-test sets and reported in Table IX for the first transcriber's data and in Table X for the second transcriber's data.

In the first row of Tables IX and X, the system is trained to distinguish between stressed (S) and unstressed (N) vocalic nuclei on the basis of the corresponding duration, amplitude, average pitch, and pitch range. The percentages refer to the stressed vocalic nuclei (S) correctly recognized, 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 following row refers to the classification subproblem $S+$ versus N. The analysis of these two fuzzy classification processes should help in understanding which input feature is the most effective in characterizing each stress class.

A similar study is reported in [24], where the effectiveness of each basic parameter to a heuristic algorithm is evaluated on the basis of the receiver operator characteristic (ROC) curve.

The fuzzy models' performances are slightly lower than the agreement percentages between the two transcribers but are comparable with the performance of other automatic algorithms [24]. The problem seems to be easier on the first transcriber's dataset, where higher discrimination percentages of stressed ($S+$ and $S-$) versus unstressed (N) syllables are obtained (Table IX compared with Table X).

The discrimination among different kinds of stress ($S+$ versus $S-$) and between minor stresses and unstressed syllables ($S-$ versus N) are much more complicated problems. In general, linguists can only reliably distinguish between fully stressed $(S+)$ and unstressed (N) syllables, whereas the distinction among different levels of stresses can not be reliably performed. The fuzzy systems' performances for this task become very low, being close to the random choice, and therefore, the corresponding performance and information gains are omitted.

In the stressed versus unstressed vocalic nuclei classification ($S+$ and $S-$ versus N), duration and amplitude produce comparable information gains for both transcribers' data sets. This means that both of them contribute circa with the same strength to the final decision process. The average pitch has the lowest information gain in both tables, which shows the low contribution of this input feature to the classification. Finally, the pitch range seems to play a more important role for the second transcriber than for the first transcriber. This agrees with the results reported in [24]. Indeed, the heuristic algorithm used in [24] produced a very good performance when using the pitch range alone, but very little improvement was obtained if combining pitch range and duration in the input vector due to an information overlapping. Moving to the $S+$ versus N problem, the fuzzy algorithm characterizes primary stress $(S+)$ by means of only amplitude

TABLE IX
INFORMATION GAINS OF THE INPUT FEATURES CHARACTERIZING STRESS IN SPOKEN AMERICAN ENGLISH FOR THE FUZZY MODEL IMPLEMENTED ON THE FIRST TRANSCRIBER'S TRAINING SETS

classification task		Information gains				% correct			
		duration	amplitude	average pitch	pitch range	S+ and S-	S+	S-	N
S (S+ and S-) vs. N	g_H	0.10	0.14	0.02	0.02	62	71	53	77
	g_G	0.13	0.17	0.02	0.02				
S+ vs. N	g_H	0.19	0.17	0.02	0.02	-	54	-	88
	g_G	0.22	0.21	0.03	0.02				

TABLE X
INFORMATION GAINS OF THE INPUT FEATURES CHARACTERIZING STRESS IN SPOKEN AMERICAN ENGLISH FOR THE FUZZY MODEL IMPLEMENTED ON THE SECOND TRANSCRIBER'S TRAINING SETS

classification task		Information gains				% correct			
		duration	amplitude	average pitch	pitch range	S+ and S-	S+	S-	N
S (S+ and S-) vs. N	g_H	0.17	0.14	0.04	0.13	56	60	40	80
	g_G	0.20	0.18	0.05	0.16				
S+ vs. N	g_H	0.22	0.11	0.09	0.08	-	53	-	83
	g_G	0.27	0.14	0.12	0.11				

and duration for both transcriber's data sets (second rows of Tables IX and X). This analysis indicates the minor role of pitch in characterizing stress, especially primary stress, in American English sentences, which agrees with what was reported in [24].

The same experiment is performed after adding the product of duration, amplitude, and average pitch to the input vector. In this case, the product is associated with the highest information gain for all the classification tasks. Even this is in agreement with the results reported in [24], where the product of these three acoustic features obtains the highest ROC curve and the best performance on the test set.

C. ECG Arrhythmia Classification

A very suitable area for fuzzy—or, in general, qualitative—decision systems consists of medical applications. 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.

In this section, we analyze an ECG classification problem that has a much higher input dimension than the previous two experiments. Because of the redundancy in the input dimension, some of the input features will present a zero or close to zero information gain. Such input features should be the ones with the lowest impact on the decision process. The goal is to investigate whether the removal of these input features influence the system's performance on the test set.

Being an almost periodic signal, the electrocardiogram (ECG) describes the electrical activity of the myocardium in time. Each time period consists of a basic waveshape, whose waves are marked with the alphabet letters P, Q, R, S, T, and U. 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. The most common types have an anomalous origin in the atria [supraventricular premature beats (SVPB)] or in the ventricula [ventricular premature beats (VPB)].

TABLE XI
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)

The MIT-BIH ECG database [25] represents a standard for 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 48 two-channel, 30-min-long records, sampled at 360 samples/s and manually annotated by trained cardiologists. QRS complexes are detected, and for each beat waveshape, a set of 12 measures [2] is extracted by using the first of the two channels in the ECG record (see Table XI).

A total of 39 records are used for this experiment, and a three-class problem (normal versus VPB's versus SVPB's) is considered. In order to produce more general results, the Jackknife procedure is applied. The selected 39 records are divided in three groups, each containing one third of the original number of records. Three different fuzzy models are constructed and

TABLE XII
INFORMATION GAINS FOR TWELVE ECG MEASURES IN A THREE-CLASS ARRHYTHMIA CLASSIFICATION PROBLEM

												% correct		
RR/RRa	QRSw	pAmp	nAmp	pQRS	nQRS	T	IVR	ST	STsl	P	PR	N	VPB	SVPB
.13	.04	.06	**.10**	.03	.02	.00	.02	.01	.02	.02	**.08**	92	80	67
.12	.03	.06	**.12**	.02	.03	-	.01	.01	.00	.02	**.08**	94	80	68
.15	.03	.07	**.15**	.06	.04	-	.02	-	.02	.01	**.09**	96	79	70
.15	.03	.06	**.14**	.05	.04	-	.02	-	.02	-	**.09**	96	79	74
.12	.03	.08	**.12**	.08	-	-	.01	-	.02	-	**.09**	95	82	72
.11	.01	.06	.06	.05	-	-	-	-	.02	-	**.09**	**95**	**83**	**72**
.08	.01	**.07**	.05	.06	-	-	-	-	-	-	**.07**	95	83	71
.07	.02	**.11**	.05	-	-	-	-	-	-	-	**.10**	95	83	70
.05	-	**.09**	.03	-	-	-	-	-	-	-	**.10**	87	70	37
.04	-	-	.03	-	-	-	-	-	-	-	**.09**	91	67	53
.05	-	-	.04	-	-	-	-	-	-	-	-	95	63	56
.06	-	-	-	-	-	-	-	-	-	-	-	96	54	00
-	-	-	.05	-	-	-	-	-	-	-	-	95	37	00
-	-	-	-	-	-	-	-	-	-	-	.10	75	39	00
-	.18	-	-	-	-	-	-	-	-	-	-	98	78	00
-	-	.03	-	-	-	-	-	-	-	-	-	93	30	00
-	-	-	-	-	-	.01	-	.01	-	-	-	89	39	00
-	-	-	-	-	-	.02	-	-	-	-	-	95	29	00
-	-	-	-	-	-	-	-	.02	-	-	-	94	36	00

tested, using two of these three groups as training set and the remaining one as test set, respectively. The output classes in each 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 [18] on each training set to discriminate the three output classes by using all 12 ECG measures. The information gains of the ECG measures and the percentages of correctly classified beats are calculated for the three fuzzy systems and reported in average in Table XII. The highest information gains are marked bold. Since the two proposed information gains assume very close values, as it could be seen in the previous experiments, only the information gain based on the entropy function is reported in Table XII.

The average performance of the three sets of fuzzy rules when all 12 ECG measures are used as input vector (see the first row of Table XII) are comparable with those reported in the literature [2]. The information gains on the left part of the row show that such performances are mainly due to the action of the prematurity degree (pd), the negative amplitude of the QRS complex, and the PR interval. The width and the positive amplitude and area of the QRS complex contribute only up to a minor extent.

Could the set of 12 ECG measures then be represented by only the input parameters with highest information gain? If the information gain of the removed input features is negligible, such reduction of the input vector should not make a big difference in terms of system's performance. In order to test this hypothesis, the ECG measures with lowest information gain are progressively removed from the input vector and the corresponding system's performance, and input features information gains are recalculated and reported in the following rows of Table XII.

At first, the T wave area (T) and the ST segment amplitude (ST) are removed from the input vector, which are the ECG measures with lowest information gain. The corresponding system's performance actually improves. Indeed, such input features were used by the system to classify outliers or exceptions in the training sets. Continuing the removal of the ECG measures with lowest information gain in the first row of Table XII, the system performance keeps improving until a maximum of 95% correctly classified normal beats—83% VPB's and 72% SVPB's (see the sixth row of Table XII)—is reached. Such a maximum in performance occurs in correspondence of an input vector with only seven components.

Performance, however, does not change much as long as the main five components of the input vector are kept: the prematurity degree, the QRS width, the positive and negative amplitude of the QRS complex, and the PR interval. These ECG measures were also the ones with a non-negligible information gain in the original analysis (see the first row of Table XII). The percentages of correctly classified beats begin to decrease dramatically only when one of these ECG measures is removed from the input vector (see the ninth row of Table XII).

The second part of the table reports the situation—information gain and system performance—of the system when using the input features with highest information gain alone. None of the five ECG measures with highest information gain in the original analysis can achieve very good performance if used alone (see the 12th–16th rows of Table XII). This was to be expected since all those features exhibit an information gain that is quite far from the maximum 1.0. The concurrence of fuzzy rules on different input dimensions seems to be necessary, particularly to recognize SVPBs. For example, the fuzzy classifier uses the positive amplitude of the QRS complex in strict connection with its duration, as we can see from the system's performance in the ninth row versus the system's performance in the tenth row of Table XII.

The last three rows of Table XII contain the information gains and the system's performance when the input vector consists of only the ECG measures with lowest information gain, namely, the T wave area and ST segment amplitude. As it was to be expected, the system's performance becomes quite poor, failing in recognizing SVPBs.

In a previous study [16] on only two files of the MIT-BIH database, the fuzzy system was retrained at each step after the removal of the input feature with lowest information gain. The information gains of the new fuzzy model resulted in more distributed across clusters of input features but, in general, was consistent with what was observed in the first experiment using all 12 ECG measures. For example, after removing an ECG measure related with the QRS morphology, the retrained system would increase the information gain of all the other ECG measures related with QRS morphology.

This investigation shows that reducing the dimension of the data set does not worsen the fuzzy system's performance if such a reduction is performed on the basis of an appropriate fuzzy feature merit measure.

VI. Conclusions

An *a posteriori* analysis of fuzzy models is presented that quantifies the influence of the input features on the decision process, that is, their discriminative power among the output classes.

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 both before and after a given input feature is used for classification. The relative difference of these two information measures defines the information gain associated with the use of this input feature, which provides a quantification of the discriminability among output classes along the analyzed input feature. This is related to the system's classification performance only if the fuzzy model is constructed on a sufficiently general set of training examples.

Artificial and real-world examples illustrated the method's potentiality. 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 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, to get insights about the implemented decision process, to look for possible errors in the decisional structure, and to compare the use of the input features by fuzzy classifiers with different performances.

Acknowledgment

The authors would like to thank W. Zong, G. Moody, and Prof. R. G. Mark from the Harvard-MIT Division of Health Sciences and Technology (Cambridge, MA) for the ECG measures and S. Greenberg of the International Computer Science Institute (Berkeley, CA) for the measures of the acoustic parameters of vocalic nuclei in spoken American English sentences. The authors would also like to thank the anonymous reviewers for their helpful feedback.

References

[1] M. Berthold and D. Hand, Eds., *Intelligent Data Analysis: An Introduction*. Berlin-Heidelberg, Germany: Springer-Verlag, 1999.

[2] W. Zong and D. Jiang, "Automated ECG rhythm analysis using fuzzy reasoning," in *Proc. Comput. Cardiol.*, 1998, pp. 69–72.

[3] Center for Spoken Language Understanding, Dept. Comput. Sci. Eng., Oregon Graduate Inst., Corvallis, "Stories corpus," Release 1.0, 1995.

[4] D. Ellis and N. Morgan, "Size matters: An empirical study of neural network training for large vocabulary continuous speech recognition," in *Proc. ICASSP*, 1999.

[5] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, pp. 81–106, 1986.

[6] ——, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

[7] C. Apte, S. J. Hong, J. R. M. Hosking, J. Lepre, E. P. D. Pednault, and B. K. Rosen, "Decomposition of heterogeneous classification problems," *Intell. Data Anal. J.*, vol. 2, no. 2, 1998.

[8] L. A. Zadeh, "A fuzzy-algorithmic approach to the definition of complex or imprecise concepts," *Int. J. Man–Mach. Stud.*, vol. 8, pp. 249–291, 1976.

[9] G. J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[10] B. M. Ayyub and M. M. Gupta, *Uncertainty Analysis in Engineering and Sciences: Fuzzy Logic, Statistics and Neural Network Approach*. Boston, MA: Kluwer, 1997.

[11] H. Bandemer and S. Gottwald, *Fuzzy Sets, Fuzzy Logic, Fuzzy Methods with Applications*. New York: Wiley, 1995.

[12] C. Z. Janikow, "Fuzzy decision trees: Issues and methods," *IEEE Trans. Syst. Man Cybern. B*, vol. Feb., pp. 1–14, 1998.

[13] A. De Luca and S. Termini, "A definition of nonprobabilistic entropy in the setting of fuzzy sets theory," *Inform. Contr.*, vol. 20, no. 4, pp. 301–312, 1972.

[14] W. J. Wang and C. H. Chiu, "Entropy and information energy for fuzzy sets," *Fuzzy Sets Syst.*, vol. 108, pp. 333–339, 1999.

[15] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.

[16] R. Silipo and M. Berthold, "Discriminative power of input features in a fuzzy model," in *Proc. Intell. Data Anal.*. New York: Springer-Verlag, 1999, pp. 85–96.

[17] M. R. Berthold and K. P. Huber, "Comparing fuzzy graphs," in *Proc. Fuzzy-Neuro Syst.*, 1998, pp. 234–240.

[18] K.-P. Huber and M. R. Berthold, "Building precise classifiers with automatic rule extraction," in *Proc. IEEE Int. Conf. Neural Networks*, vol. 3, 1995, pp. 1263–1268.

[19] R. A. Fisher, "The use of multiple measurements in taxonomic problems," in *Annual Eugenics, II*. New York: Wiley, 1950, vol. 7, pp. 179–188.

[20] C. Blake, E. Keogh, and C. J. Merz, "UCI repository of machine learning databases," Dept. Inform. Comput. Sci., Univ. Calif., Irvine, http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.

[21] D. van Kuijk and L. Boves, "Acoustic characteristics of lexical stress in continuous telephone speech," *Speech Commun.*, vol. 27, pp. 95–111, 1999.

[22] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 469–481, 1994.

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

[24] R. Silipo and S. Greenberg, "Automatic transcription of prosodic stress for spontaneous English discourse," in *Proc. XIVth Int. Congr. Phonetic Sci. (ICPhS)*, vol. 3, 1999, p. 2351.

[25] MIT-BIH Database Distributor, Beth Israel Hospital, Biomed. Eng., Div., KB-26, Boston, MA.

Rosaria Silipo (M'98) received the electrical engineering degree from the University of Florence, Florence, Italy, in 1992. She developed her doctorate thesis at the University of Florence about the automatic analysis of the ECG signal and received the Ph.D. degree in bioengineering from the Politecnico di Milano, Milan, Italy, in 1996.

In 1994, she was with the Massachusetts Institute of Technology, Cambridge, as Visiting Researcher. From 1996 to 1997, she was with Siemens AG, Munich, Germany, as Visiting Scientist, where she worked on the nonlinear analysis of biomedical time series. Until April 2000, she was a Post-Doctoral Fellow with the International Computer Science Institute, University of California, Berkeley. Her current research and interests include speech recognition and, particularly, the analysis of prosodic features in speech.

Michael R. Berthold (M'94) received the Ph.D. degree in computer science from the University of Karlsruhe, Karlsruhe, Germany.

From Fall 1997 until early 2000, he was a Research Fellow with the Berkeley Initiative in Soft Computing (BISC) and a Lecturer at the University of California, Berkeley. He was a Visiting Researcher at Carnegie Mellon University, Pittsburgh, PA, from 1991 to 1992 and at Sydney University, Sydney, Australia, in 1994. He was a Research Engineer with Intel Corporation, Santa Clara, CA, in 1993. His research interests include neural networks, fuzzy logic, and intelligent data analysis. He edited—together with D. J. Hand—the textbook *Intelligent Data Analysis: An Introduction* (New York: Springer-Verlag).

Dr. Berthold is a Member of the Steering Committee of the IDA conference series.