Combination of Classifiers using the Fuzzy Integral for Uncertainty Identification and Subject Specific Optimization

Application to Brain-Computer Interface

Francesco Cavrini[1], Lucia Rita Quitadamo[2], Luigi Bianchi[3] and Giovanni Saggio[2]

[1]*Department of Computer, Control, and Management Engineering, University of Rome "La Sapienza", Rome, Italy*

[2]*Department of Electronic Engineering, University of Rome "Tor Vergata", Rome, Italy*

[3]*Department of Civil Engineering and Computer Science Engineering, University of Rome "Tor Vergata", Rome, Italy*

Keywords: Brain-Computer Interface (BCI), Combination of Classifiers, Fuzzy Integral, Fuzzy Measure, Multi-Classifier Systems (MCS).

Abstract: In this paper we propose a framework for combination of classifiers using fuzzy measures and integrals that aims at providing researchers and practitioners with a simple and structured approach to deal with two issues that often arise in many pattern recognition applications: (*i*) the need for an automatic and user-specific selection of the best performing classifier or, better, ensemble of classifiers, out of the available ones; (*ii*) the need for uncertainty identification which should result in an abstention rather than an unreliable decision. We evaluate the framework within the context of Brain-Computer Interface, a field in which abstention and inter-subject variability have a remarkable impact. Analysis of experimental data relative to five subjects shows that the proposed system is able to answer such needs.

1 INTRODUCTION

Multi-Classifier Systems (MCSs), one of the key areas of current machine learning research, constitute a vast family of pattern recognition techniques which have proved to be useful in increasing the overall classification accuracy and robustness; such approaches are known in the literature with a plethora of terms, e.g. classifier fusion, ensembles of learning machines, combination of (multiple) classifiers, ensemble methods, mixture of experts (Kuncheva, 2001; Ranawana and Palade, 2006). Traditionally, MCS have been viewed as a means for improving classification accuracy and reducing its variance. In this paper, we propose the use of classifier combination in a different fashion. Our study is motivated by two issues that arise in many pattern recognition research and application fields, e.g. human-machine interfaces:

i. There is often no evidence of a single classifier outperforming all the others for all the users of the system.

ii. Misclassification is more dangerous or has a greater impact on performance and usability than abstention.

The paper is concerned with the development of a framework for combination of classifiers that can help in dealing with issue *i* and *ii* and that can be applied to a variety of systems with minimal effort and no changes to their structure. We feel that:

- The use of multiple approaches and the automatic, user-specific, selection of those that perform best could be a step towards the realization machine learning infrastructures ready to be used by different subjects.
- As it integrates decisions from different sources, combination of classifiers is promising of being better at uncertainty identification than a single pattern recognition technique.

We evaluate the proposed framework within the context of Brain-Computer Interface (BCI), a field in which issue *i* and *ii* are of particular interest (see section 5 for further information).

The rest of the paper is organized as follows. In section 2 we introduce the basic principles and the structure of a generic classifier combination system. Section 3 is devoted to the presentation of the theoretical concepts on which the proposed strategy is grounded. In section 4 and 5 are illustrated, respec-

Cavrini F., Quitadamo L., Bianchi L. and Saggio G..
Combination of Classifiers using the Fuzzy Integral for Uncertainty Identification and Subject Specific Optimization - Application to Brain-Computer Interface.
DOI: 10.5220/0005035900140024
In *Proceedings of the International Conference on Fuzzy Computation Theory and Applications* (FCTA-2014), pages 14-24
ISBN: 978-989-758-053-6

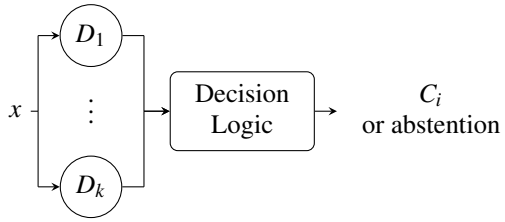


Figure 1: Logical schema of a classifier combination system.

tively, the proposed framework and its application to BCI. The results obtained in the offline analysis of data from five subjects are presented in section 6, which is followed by a discussion of experimental findings and practical implementation issues. Finally, we conclude and remark on future work.

2 FUNDAMENTALS OF COMBINATION OF CLASSIFIERS

Let $C_1,\ldots,C_n$ be the n classes of a pattern recognition task T and let $D_1,\ldots,D_k$ be k classifiers for T. In this paper, with *combination of classifiers system* (Figure 1), we denote a MCS that outputs the class C_i to which an input feature vector is expected to belong on the basis of the classification performed by the D_j ($j = 1,\ldots,k$) only. Moreover, the system may abstain if some predefined criteria are not met, e.g. more than half of the classifiers shall be in agreement. Throughout the paper we will often refer to the D_j as *first level classifiers*.

The aggregation of first level classifiers output depends on the information they provide. If a classifier returns only the label of the chosen class, then combination typically reduces to some form of *voting*; instead, if a classifier assigns to each class a value representing the extent to which it believes the input vector belongs to that class, then more sophisticated techniques, such as *weighted averaging* or *fuzzy integrals*, can be used (Kuncheva, 2001).

In order for the combination to be successful, the first level classifiers should be *different* (Ranawana and Palade, 2006). There is no general agreement in the scientific community about the definition of the concept of *diversity among classifiers*, here we say that two classifiers are different if at least one of the following holds:

- They belong to different pattern recognition approaches, e.g. one is a Support Vector Machine and the other is an Artificial Neural Network.
- They work in different feature spaces.
- Even if they belong to the same pattern recognition approach, they have been configured differently.
- They have been trained on disjoint subsets of the available data.

More formal definitions and measures of diversity in classifier ensembles lie beyond the scope of this paper and the interested reader is referred to (Ranawana and Palade, 2006). The terminology introduced in this paragraph will be used hereafter.

3 THEORETICAL BACKGROUND

Given our perspective, the following theoretical introduction will be limited to finite spaces; for an extensive coverage of fuzzy measures and integrals see (Grabisch et al., 1995) and references therein. We adopt the following notation and conventions: $\emptyset$ denotes the empty set; $|X|$ indicates the cardinality of set X, and $\mathcal{P}(X)$ denotes its power set; $0! = 1$, as usual.

3.1 Fuzzy Measure and Integral

Given a system with n inputs $x_1,\ldots,x_n$, a typical way to express the worth of each input and of each possible coalition of inputs with respect to the overall output of the system is to define a *measure* on $X = \{x_1,\ldots,x_n\}$. However, the additivity of the measure often turns out to be quite restrictive, as it does not allow us to model all those scenarios in which the sources manifest positive or negative synergy when put together into a coalition. As a solution to that rigidness, Sugeno introduced the concept of *fuzzy measure* (Sugeno, 1974).

Definition 1. *Let X be a finite set. A fuzzy measure μ on X is a set function defined on $\mathcal{P}(X)$ satisfying the following axioms:*

i. $\mu(\emptyset) = 0$ *(vanishes at the empty set).*

ii. $\forall A,B \in X.\ A \subseteq B \Rightarrow \mu(A) \leq \mu(B)$ *(monotonicity).*

If, in addition, $\mu(X) = 1$ then the fuzzy measure is said to be normalized.

Having extended the measure we also need to extend the *integral*. Various definitions of integral operators with respect to fuzzy measures have been proposed (Murofushi and Sugeno, 2000), the most used in practical applications being the so called *Sugeno integral* (Sugeno, 1974) and *Choquet integral*. Among these, the one originally due to Choquet (Choquet, 1953) is the only proper generalization of the ordinary integral, i.e. the Lebesgue integral is recovered

whenever the fuzzy measure is additive. For a characterization and discussion of the two integrals we refer the reader to (Grabisch et al., 1995; De Campos and Jorge, 1992; Grabisch, 1996). The Choquet integral is the one of choice in this study because of its mathematical properties, both from a theoretical point of view and from a practical perspective. In fact, as it will be shown in section 4.2, it allows us to express the learning task in terms a convex quadratic program, which can be solved efficiently by means of well known algorithms.

Definition 2. *Let X be a finite set of n elements, and let μ be a fuzzy measure on X. Let $f : X \to \mathbb{R}^+$. Permute the elements of X so that $f(x_1) \leq \cdots \leq f(x_n)$ where x_1 is the first element of X permuted, x_2 the second, and so on. The Choquet integral of f with respect to μ is defined as:*

$$\sum_{i=1}^{n} (f(x_i) - f(x_{i-1}))\,\mu(A_i) \tag{1}$$

where $f(x_0) = 0$ and $A_i = \{x \in X \,|\, f(x) \geq f(x_i)\}$.

3.2 Shapley Value and Interaction Index

Once we have a fuzzy measure on the set of available information sources, it would be interesting to estimate the contribution that each of them brings to the task at hand; for such a purpose the *Shapley value* (Shapley, 1953) can be used.

Definition 3. *Let $X = \{x_1, \ldots, x_n\}$ be a finite set, and μ be a fuzzy measure on X. The Shapley value, or importance index, of element x_i with respect to μ is defined as*

$$\sum_{A \subseteq X \setminus \{x_i\}} \frac{(n - |A| - 1)!\,|A|!}{n!} \Delta_{x_i}(A) \tag{2}$$

where $\Delta_{x_i}(A) = \mu(A \cup \{x_i\}) - \mu(A)$.

Although the Shapley value provides us with precious information about the importance of each source, it does not give us clues about their pairwise interaction. With that objective, Murofushi and Soneda introduced the *interaction index* (Murofushi and Soneda, 1993). Given two sources x_i and x_j, the interaction index, I_{x_i,x_j}, is such that:

- If the two sources manifest a positive synergy when working together, then $I_{x_i,x_j} > 0$.
- If the two sources hamper each other, then $I_{x_i,x_j} < 0$.
- If the two sources do not gain neither lose anything from being together, then $I_{x_i,x_j} = 0$.

Grabisch extended the interaction index to coalitions of any number of elements (Grabisch, 1997).

Definition 4. *Let $X = \{x_1, \ldots, x_n\}$ be a finite set, and μ be a fuzzy measure on X. The extended (or generalized) interaction index, I_S, of the coalition $S \subseteq X$ with respect to μ is defined as:*

$$\sum_{A \subseteq X \setminus S} \frac{(n - |A| - |S|)!\,|A|!}{(n - |S| + 1)!} \sum_{B \subseteq S} (-1)^{|S|-|B|} \mu(A \cup B) \tag{3}$$

It can be shown (Grabisch, 1997) that the extended interaction index is a proper generalization of the aforementioned concepts of importance and interaction, i.e. the Shapley value and the pairwise interaction index are recovered whenever the coalition is made up of, respectively, one or two elements.

4 PROPOSED FRAMEWORK

In the following we assume that each of the first level classifiers, for each feature vector in input, provides a vector of n values whose i-th entry indicates the "score" assigned to class i, the highest value being the one corresponding to the class the learner believes the input vector belongs to. Note that such an assumption is not much restrictive, as many of the most widely used classifiers readily provide such information, e.g. for a Bayesian classifier the "scores" are the *a-posteriori* probabilities.

4.1 Overview

Figure 2 depicts the structure of the proposed framework. From the k available classifiers $D_1, \ldots, D_k$, we build n class-specific logical ensembles $E_1, \ldots, E_n$ of s classifiers each ($1 \leq s \leq k$). The coalition E_i will consist of those classifiers that best cooperate for recognition of inputs belonging to class C_i. In section 4.4 we describe how to build the logical ensembles from training data in a completely automatic way.

As each of the first level classifiers assigns scores to classes in its own way (e.g. for a Bayesian classifier the scores are probabilities, whereas for a SVM they could be distances), direct combination of the information they provide is not legitimate. We need a transformation procedure (DRI box in Figure 2) to map values from the output domain of each classifier into a common, classifier-independent, space. For each classifier D_j^i of each logical ensemble E_i, we

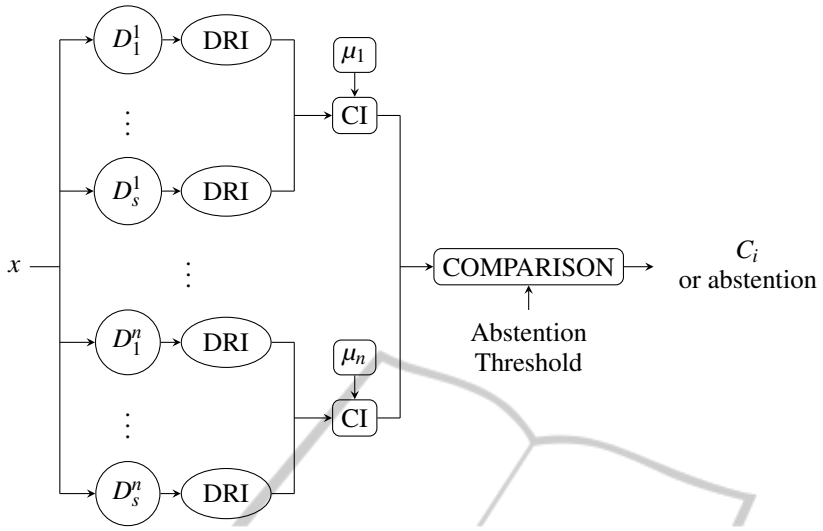


Figure 2: Logical schema of the proposed framework for class specific combination of classifiers. The flow of information and the meaning of the blocks is described in detail in section 4.1.

compute a value, DRI_j^i, where *DRI* stands for *Decision Reliability Index*, that can be regarded as the degree of belief in the proposition *"The input vector x belongs to class C_i"*, according to classifier D_j^i and depending on the reliability of its decision. That value lies in the unit interval and has the following interpretation: $DRI_j^i = 1$ indicates absolute certainty that the input belongs to class C_i; $DRI_j^i = 0$ indicates absolute certainty that the input does not belong to class C_i; $DRI_j^i \in (0,1)$ expresses an intermediate degree of belief between the two aforementioned extremes. Further details on the transformation are given in section 4.3.

For each logical ensemble E_i, let f_i be the function that assigns to each classifier in the ensemble its decision reliability index for class C_i, i.e. $f_i\left(D_j^i\right) = DRI_j^i$ ($j = 1,\ldots,s$). We integrate f_i using the Choquet integral (CI box in Figure 2) with respect to the fuzzy measure defined on E_i, i.e. μ_i. All fuzzy measures are learned from data in the training phase (section 4.2). Let p_i denote the result of the integration. The response of the framework is either the class C_i having the maximum p_i ($i = 1,\ldots,n$) or an abstention. In particular, our policy for rejection is as follows. Let p_{first} and p_{second} denote, respectively, the first and the second maximum value of p_i. Let τ be an abstention threshold. If $p_{first} - p_{second} \leq \tau$, then it is safer to abstain rather than providing a not enough reliable decision.

4.2 Training

The ensemble specific fuzzy measures $\mu_1,\ldots,\mu_n$ are learned from data by means of an approach grounded on least squares optimization. In particular, for each class C_i, the error criterion to minimize is:

$$\sum_{t=1}^{T}\left(y_i^t - CI_{\mu_i}\left(f_i^t\right)\right)^2 \tag{4a}$$

subject to:

$$\mu_i(A) \leq \mu_i(B) \quad \text{whenever } A \subseteq B \subseteq E_i \tag{4b}$$

where:

- T is the number of trials.
- $CI_{\mu_i}\left(f_i^t\right)$ indicates the Choquet integral of f_i^t with respect to μ_i.
- y_i^t represents the desired output for trial t and is equal to the maximum (resp. minimum) value that the Choquet integral can assume if trial t belongs (resp. does not belong) to class C_i.

Due to the peculiarities of the Choquet integral, it is possible to express the constrained least square optimization problem in terms of a convex quadratic program, which is easier and more efficient to solve. See (Grabisch et al., 1995) for further details.

A common issue in supervised learning algorithms is the minimum amount of training data required to involve all model coefficients. For fuzzy measure identification, we have the following lower bound (Grabisch et al., 1995), where s denotes the number of elements of the set on which we want to define the fuzzy measure, i.e. the size of the class-specif ensembles of classifiers in our framework:

$$T_{min} = \frac{s!}{\left(\frac{s}{2}\right)!\left(\frac{s}{2}\right)!} \tag{5a}$$

if s is even;

$$T_{min} = \frac{s!}{\left(\frac{s-1}{2}\right)!\left(\frac{s+1}{2}\right)!} \tag{5b}$$

if s is odd. Note that if $T \geq T_{min}$ then it is not guaranteed that all fuzzy measure coefficients will be involved, but if $T < T_{min}$ some of them will certainly not be used.

4.3 Decision Reliability Index

The transformation of classifier outputs into decision reliability indexes (DRIs) involves two consecutive steps: firstly a linear normalization and afterwards a non-linear mapping into the degree of belief space. Let $\boldsymbol{d}$ be the n-dimensional vector containing the output of a first-level classifier, with its i-th entry indicating the score the classifier assigned (for the feature vector in input) to the i-th class. We linearly project $\boldsymbol{d}$ into $[-1,1]^n$ so to obtain a new vector, $\boldsymbol{d}_\pi$, that lies in a space that is independent from the output domain of the classifier. Next, we nonlinearly map $\boldsymbol{d}_\pi$ into the degree of belief space using a sigmoid function:

$$sig\,(x) = \frac{1}{1+e^{-a(x-c)}} \tag{6}$$

where $a \in [0,\infty)$ is the *slope factor* and $c \in [-1,1]$ is the *crossover point*.
For each $\boldsymbol{d}_\pi$ we estimate a and c by means of nonlinear optimization. The objective function is given by the following consideration: since the input vector has to belong to one of the classes of the pattern recognition task, than the sum of the DRIs shall equal certainty, i.e.

$$\|sig\,(\boldsymbol{d}_\pi)\|_1 = 1 \tag{7}$$

In addition, we impose that:

i. c shall lie in the interval given by the first and the second maximum value of $\boldsymbol{d}_\pi$.

ii. a shall be upper-bounded.

These constraints arise to penalize decisions taken with considerable uncertainty and reward those that instead reflect good discrimination by the classifier. Firstly, it should be noted that as a increases the sigmoid function tends to 1 if $x > c$, to $1/2$ if $x = c$, and to 0 if $x < c$. Such an extreme behavior should be avoided as, with the aforementioned boundary conditions on c, it impairs the DRI significance: we would assign complete certainty to the class chosen by the classifier, neglect of the presence of the others and of the uncertainty hidden in every decision. By graphic inspection we have found that a reasonable upper bound for a is 20. To realize how the constraints penalize uncertain decisions, consider the case in which two classes appear both probable for the input vector to belong to; in such a situation the classifier would assign a high and similar measurement to both of them. It is easy to see that, limiting c and a as previously specified, also the DRIs relative to those classes will be similar and approximately around $1/2$. Moreover, the DRIs relative to the other classes will probably be not negligible and therefore, to enforce (7), we will have to lower the DRIs relative to the overall, uncertain decision.

4.4 Classifier Selection

To identify the logical ensemble of classifiers E_i that is best at recognition of trials belonging to class C_i, we rely on the following intuitive observation: a good team is composed of players being themselves good players and that positively collaborate towards the achievement of a common goal. Between those two not necessary correlated criteria, i.e. individual skill and group interaction, we believe the latter is the one that influences the strength of an ensemble most. Such considerations led us to the following classifier selection strategy (Algorithm 1): initially take the best classifier and then incrementally grow the ensemble by including the classifier that best interacts with those already in. We use the generalized interaction index (3) to estimate synergy among members of a coalition. Recall that such an index is a proper generalization of the Shapley value (2), thus the first classifier selected will be the one with the highest Shapley value, i.e. the most important in terms of contribution to the pattern recognition process.

4.5 Abstention Threshold Selection

Typically, there exists a trade-off between misclassification and abstention. If we increase the abstention threshold, then we reduce the number of misclassifications to the detriment of the amount of abstentions, and vice-versa. Depending on the application, we may assign to each error a penalty value w_e, which represents the cost of misclassification from the point of view of system performance, usability, safety. Similarly, we may assign to each abstention a penalty value w_a. In general, $w_a \leq w_e$. In such a model, given a training dataset, the subject-specific optimal abstention threshold is the one that minimizes:

$$\sum_{t=1}^{T} R_\tau(x_t) \tag{8a}$$

Algorithm 1: Classifier selection.

Data: Class C_i-specific fuzzy measure μ_i on the set $D = \{D_1, \ldots, D_k\}$ of all available classifiers; μ_i is learned from data in a preliminary training phase. The number s of classifiers to select.

Result: Class C_i-specific logical ensemble E_i of classifiers.

$E_i = \emptyset$;
while $|E_i| < s$ **do**
  foreach $D_j \in D$ **do**
    $I_j =$ generalized interaction index of the ensemble $E_i \cup \{D_j\}$;
  end
  $D_{best} = \underset{D_j \in D}{\operatorname{argmax}} I_j$;
  $E_i = E_i \cup \{D_{best}\}$;
end

with:

$$R_\tau(x_t) = \begin{cases} 0 & \text{if } \widehat{y}_t(\tau) = y_t \\ w_a & \text{if } \widehat{y}_t(\tau) = \text{abstention} \\ w_e & \text{otherwise} \end{cases} \quad (8b)$$

subject to:

$$0 \leq \tau < 1 \quad (8c)$$

where:

- T is the number of trials.
- x_t is the t-th input feature vector.
- $\widehat{y}_t(\tau)$ is the response of the framework, using abstention threshold τ, to x_t.
- y_t is the desired response.

Even though there are techniques for solving non-linear programs as the one above, given the fact that the interval for the only free variable τ is limited and extreme precision is not fundamental, we suggest to pursue an approximate solution using *grid-search*, which is much easier to implement and faster to execute.

5 APPLICATION TO BCI

A Brain-Computer Interface (BCI) system is an Assistive Technology device that allows to translate brain activity into commands towards an output peripheral (Wolpaw et al., 2002). It is mainly intended for severely disabled people who, after traumas or neurodegenerative diseases (e.g. amyotrophic lateral sclerosis), have lost control of their muscles and therefore any possibility to communicate towards the external (Sellers et al., 2010; Hochberg et al., 2012). A BCI system records brain activity by means of sensors, the most diffuse technique being the electroencephalography (EEG), and translates signal variations (originating from the execution of a mental task) into an output command that can be fed into different devices, e.g. a spelling interface, a cursor on a screen, a wheelchair, a robotic hand.

The classification of mental states is a crucial step in the BCI chain. First of all, despite the remarkable number of articles related to this issue, an optimal classifier, that could be adapted in the most performing way to different subjects, has not yet emerged. Moreover, an accurate detection process is fundamental for the whole BCI system, especially when the final application is not just devoted to a simple communication task, by means, e.g., of a spelling interface (Krusienski et al., 2008), but when the BCI pilots a wheelchair (Rebsamen et al., 2010) or a robotic hand (Muller-Putz and Pfurtscheller, 2008).

5.1 Ensemble of Classifiers and Abstention in BCI

Ensemble of classifiers were already used in BCI, and proved to be among the most powerful classification techniques. For example, in (Rakotomamonjy and Guigue, 2008) the authors considered an ensemble of SVMs to classify data from the BCI Competition III (http://www.bbci.de/competition/iii/), achieving very high accuracy at the cost of a huge amount of required training data. In (Faradji et al., 2008), instead, the authors used an ensemble of radial basis function neural networks for recognition of intentional control/no control states in order to activate a self-paced BCI switch; whereas in (Johnson and Krusienski, 2009) an ensemble of SWLDAs was the classification choice for a P300 speller BCI. Despite the high accuracies achieved, all the mentioned studies involve ensembles of classifiers of the same kind and do not take into account the advantages of abstention-capable strategies. A classification method that considers the abstention alternative was implemented in (Schettini et al., 2014) and (Aloise et al., 2013), where it was stated that a closer to reality evaluation of BCI systems should include the contribute of abstentions among the performance assessment criteria.

5.2 The P300-based Speller

One of the most diffused protocol in BCI research is the one based on the so called *matrix speller*, or *P300 speller* (Farwell and Donchin, 1988). The sub-

ject faces a computer screen that displays a matrix of alphabet letters and other symbols, e.g. single digit numbers, space and undo commands. In a trial, each row and column flash randomly for F times; each flash or stimulus lasts N ms and there is an Inter-Stimulus Interval of ISI ms. The subject is asked to count how many times the symbol he/she wants to communicate flashes. Each flash of the desired symbol elicits the *P300 component* of an *event-related potential* and it is therefore possible to infer from the registered brain electrical activity which character or command the user was focusing on.

5.3 Experimental Setup

We evaluated the proposed framework within the context of a P300 speller. An EBNeuro Mizar System (Florence, Italy) was used for EEG recording. Signal preprocessing and first level classification were performed by means of the NPXLab Suite (Bianchi et al., 2009), whereas the proposed framework was implemented as a set of dynamic libraries in the C programming language. The settings of the matrix speller protocol we refer to are quite standard and have been already used in (Bianchi et al., 2010): $F = 15$, $N = 100$ ms and $ISI = 80$ ms. The EEG activity was recorded using 61 sensors positioned according to the *10-10 international system*, at a sampling rate of 256 Hz, with reference electrode positioned between AFz and Fz and ground between Pz and POz. After acquisition, data was band-pass filtered between 0.5 and 30 Hz and artifact (e.g. eye-blinks) removal was performed by an expert neurophysiopathology technician. Six of the most used classifiers in the matrix speller protocol (Krusienski et al., 2006) were considered: Bayesian classifier, Artificial Neural Network (ANN), Shrunken Regularized Linear Discriminant Analysis (SRLDA), Stepwise Linear Discriminant Analysis (SWLDA), Support Vector Machine with linear kernel (SVM-LIN) and with radial basis function kernel (SVM-RBF). The size of the class-specific ensembles was limited to 4 classifiers to avoid excessive computational complexity.

6 RESULTS

Five healthy subjects (3 men and 2 women, aged from 22 to 43 years) participated in the experiments. For each subject, 6 sessions were recorded. A small break separated two consecutive sessions and each of them was concerned with the communication of 6 different symbols. The first level classifiers and the framework were trained, respectively, on the first 12 characters and on data from the third session. Testing involved the last 18 symbols.

To evaluate performance we introduce the notion of *weighted accuracy*:

$$WA = 1 - ER - 0.5 \times AR \tag{9}$$

where ER is the error-rate and AR is the abstention-rate. In the weighted accuracy, errors are assigned a penalization factor that is double of that of abstentions, this is because correcting a wrongly classified symbol requires (correct) recognition of the "undo" command and re-communication of the desired one, whereas an abstention induces the need to perform only the latter of these two steps.

Table 1 shows the weighted accuracy achieved, on the test set, by the first level classifiers and by the framework. To facilitate visualization of relationships between the classifiers and the framework, the same data is also reported in Figure 3.

Firstly, it is possible to notice that the same classifier does not perform equally well for all the subjects. For example, the artificial neural network, which is quite good for subject A, D and E, performs poorly for subject B. Even the Bayesian classifier, which is the best of the available ones for most of the users, is surpassed by SVM-LIN when considering the fourth subject.

Secondly, the proposed framework leads to a weighted accuracy that, for each subject, is similar to or higher than that of the best of the available classifiers. To further investigate on this, Table 1 also reports the percentage improvement achieved by the framework with respect to the average and the best of the first level classifiers.

Finally, by direct look at the confusion matrices (which we do not report for the sake of brevity) we could note that the improvement the proposed framework leads to often comes from its ability to identify uncertain situations and turn them from misclassification into abstentions, thus making the overall system safer and more pleasant to use.

7 DISCUSSION

Experimental evidence shows that the proposed framework is able to deal with the two issues that motivated our study. In particular, it reaches a level of performance similar to or greater than that of the best first level classifier, which, nevertheless, is not the same for all subjects. Hence, the framework eliminates the need for a preliminary configuration phase in which an expert has to either find a classifier that performs well for most of the subjects or select the

Table 1: Weighted accuracy, as defined in (9), of the first level classifiers and of the proposed framework; percentage improvement achived by the framework with respect to the average (column I_{avg}) and the best (column I_{best}) of the first level classifiers. *Bayesian* denotes the Bayesian classifier; *ANN* stands for Artificial Neural Network; *SRLDA* represents Shrunken Regularized Linear Discriminant Analysis; *SWLDA* indicates StepWise Linear Discriminant Analysis; *SVM-LIN* and *SVM-RBF* denote Support Vector Machine with, respectively, linear and radial basis function kernel.

Subject	Bayesian	ANN	SRLDA	SWLDA	SVM-LIN	SVM-RBF	Framework	I_{avg}	I_{best}
A	0.863	0.853	0.803	0.823	0.858	0.848	0.868	3.2	0.58
B	0.782	0.651	0.590	0.697	0.641	0.742	0.767	12.17	-1.93
C	0.729	0.681	0.688	0.688	0.737	0.659	0.725	4.07	-1.50
D	0.757	0.747	0.722	0.681	0.762	0.742	0.790	7.43	3.64
E	0.919	0.904	0.813	0.909	0.909	0.893	0.939	5.38	2.19

best one in a subject-specific manner. This can help in building the pattern recognition system for the task at hand: one could simply use all the algorithms the literature suggests, or the ones he/she has at disposal, and then let the framework perform a subject and class specific adaption that would probably lead to an optimal level of performance. Moreover, by taking into account the output of class-specific ensembles of classifiers, in many situations the framework is better at uncertainty identification than a single classifier alone. Vague decisions are often turned from misclassification into abstentions, a property that is of particular importance in safety-critical applications.

7.1 Classifier Selection, Training and Computation Complexity

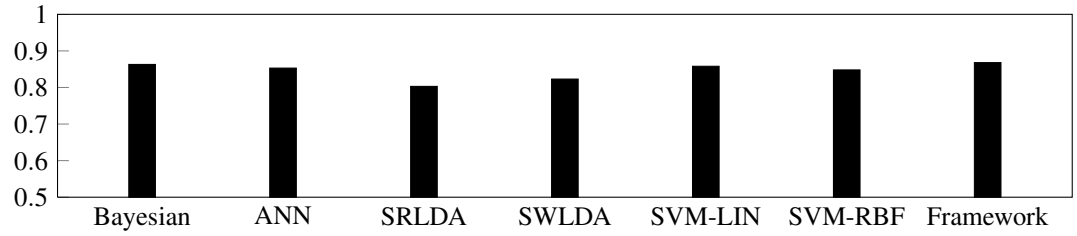
Throughout the paper, we have assumed the need to build ensembles of s classifiers out of the k available ($1 \leq s \leq k$). Nevertheless, it would be reasonable to ask oneself: if we have at our disposal k classifiers, why do we not use them all? The reason is twofold: firstly, richness in information comes at the cost of increased computational complexity and, secondly, the relationship between input information and discriminatory capability is, roughly speaking, influenced more by the quality than the quantity of information. Those issues are not specific to the proposed framework but rather typical of many pattern recognition approaches. In fact, building those ensembles can be regarded as a particular case of the well-known problem of *feature selection*. In addition, a relevant issue is the trade-off between the number s of classifiers to select and computational tractability of the combination process. As a fuzzy measure on a set of s elements requires $2^s - 2$ coefficients to be defined, we have to deal with an exponential number of variables in the training procedure and therefore s should be limited to a small value, e.g. a reasonable empirical bound is 8.

Given the previous observations, it may seem contradictory that in the proposed classifier selection procedure (Algorithm 1) we use the fuzzy measure μ_i on the entire set of available classifiers. There is no inconsistency in that, and to explain why we need to introduce the notion of *k-additive fuzzy measure* (Grabisch, 1997).

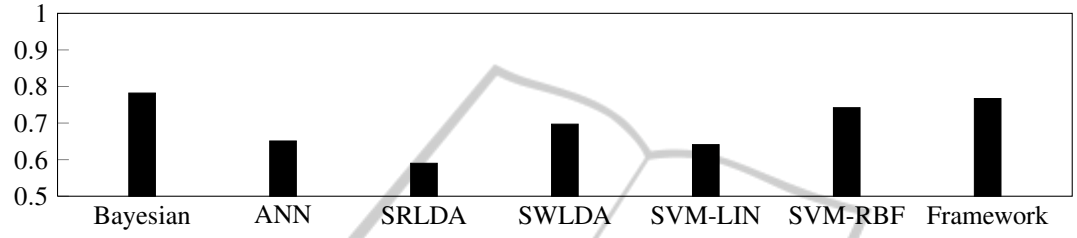
A k-additive fuzzy measure combines the power of the fuzzy measure with the simplicity of the ordinary measure, thus resulting in a good trade-off between expressiveness and computational tractability. A k-additive fuzzy measure limits interaction to subsets of cardinality $\leq k$, and the values of the fuzzy measure for the remaining subsets are completely predetermined by the additivity constraints. It follows that to define a k-additive fuzzy measure on a set of n elements, we do not need $2^n - 2$ coefficients but only $\sum_{i=1}^{k} \binom{n}{i}$. The process of learning a k-additive fuzzy measure from data is similar to that presented in section 4.2, see (Miranda and Grabisch, 1999) for further details.

Since in the classifier selection algorithm we use the fuzzy measure μ_i on the entire set of available classifiers only to estimate interaction between up to s classifiers, we do not actually need the full power of a fuzzy measure, a s-additive fuzzy measure is sufficient. This makes the proposed approach computationally feasible in most of the practical situations, e.g. for $s = 4$ and 10 available classifiers, the number of parameters to identify is 385, instead of the 1022 that characterize the corresponding fuzzy measure.

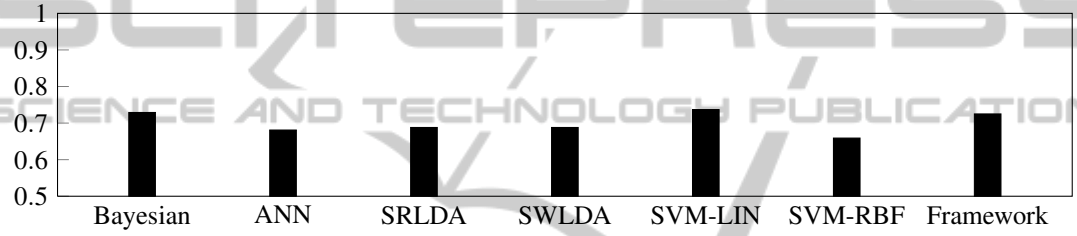
Even though we have not provided a formal analysis of space and time requirements of the proposed approach, it should be clear that the most burdensome procedures are those related to fuzzy measure learning. Notwithstanding, with classifier selection it is possible to limit the exponent to a small value, e.g. 4 in our matrix speller application, and thereby the aforementioned complexity will not significantly affect performance. For example, each of the test described in the results section took about a second on a laptop running Windows 7 with an Intel Core i5 CPU (2.4 GHz) and 4 GB of RAM. In addition, it should be noted that combination of classifiers by means of



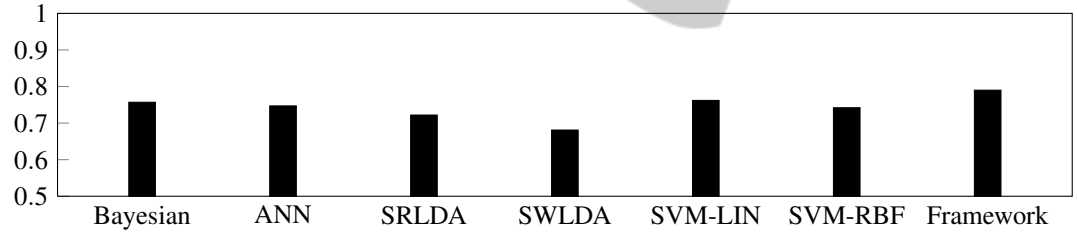
(a) Subject A.



(b) Subject B.



(c) Subject C.



(d) Subject D.

(e) Subject E.

Figure 3: Weighted accuracy, as defined in (9), of the first level classifiers and of the proposed framework for each subject. *Bayesian* denotes the Bayesian classifier; *ANN* stands for Artificial Neural Network; *SRLDA* represents Shrunken Regularized Linear Discriminant Analysis; *SWLDA* indicates StepWise Linear Discriminant Analysis; *SVM-LIN* and *SVM-RBF* denote Support Vector Machine with, respectively, linear and radial basis function kernel.

the proposed framework is highly parallelizable. Obviously, first level classifiers can operate in parallel; moreover, since each class-specific ensemble is independent of the others, the implementation of the proposed method can be almost entirely parallelized. Finally, once the class-specific ensembles have been built and the fuzzy measures have been learned, classification of a new trial requires a negligible compu-

tational time and therefore we argue that the proposed approach is suitable for real-time application too.

8 CONCLUSION

Our study has been motivated by two issues that arise in many pattern recognition applications:

i. There is often no evidence of a single classifier outperforming all the others for all the users of the system.

ii. Misclassification is more dangerous or has a greater impact on performance and usability than abstention.

To address such issues we have proposed a framework for combination of classifiers that is able to:

- Automatically select the best performing ensemble of classifiers for each subject and each class of the problem.
- Better identify vague situations by taking advantage of the information provided by many different sources, instead of a single one.

The framework is based on a general paradigm of information fusion by means of fuzzy measures and integrals (Kuncheva, 2001; Grabisch et al., 1995) and presents novel solutions for what concerns the overall architecture, the process of classifier selection and the normalization of their output. Moreover, it is applicable as a "black-box" to any domain, without the need to change or adapt the pattern recognition system the experimenter has set up, a feature which we feel is important in order to speed up the process of constructing a valid configuration for the problem of interest.

We have performed a preliminary validation of the proposed method within the context of a P300-based matrix speller Brain-Computer Interface. Even though only a restricted number of subjects participated in the experiments, we were nevertheless able to point out the importance of issue *i* and *ii* and the prompt response of the framework. Results show that the proposed method is able to reach, for each subject, a level of performance significantly greater than the average of the available classifiers and similar to or greater than that of the best one.

To further validate the proposed approach, more tests are needed, and this is part of our future work. We would like to apply the framework into different contexts, to confirm the positive outcomes obtained in this study and/or evidence possible drawbacks. Moreover, we are interested in comparing the proposed approach with other popular ensemble methods, e.g. Boosting, Mixture of Experts, Error-Correcting Output Codes, Stacking (Alpaydin, 2009). Finally, we would like to compare the proposed classifier selection algorithm to the ones already present in the literature.

REFERENCES

Aloise, F., Aricò, P., Schettini, F., Salinari, S., Mattia, D., and Cincotti, F. (2013). Asynchronous gaze-independent event-related potential-based brain–computer interface. *Artificial intelligence in medicine*, 59(2):61–69.

Alpaydin, E. (2009). *Introduction to Machine Learning*. The MIT Press, 2nd edition.

Bianchi, L., Quitadamo, L. R., Abbafati, M., Marciani, M. G., and Saggio, G. (2009). Introducing NPXLab 2010: a tool for the analysis and optimization of P300 based brain–computer interfaces. In *2nd International Symposium on Applied Sciences in Biomedical and Communication Technologies*, pages 1–4. IEEE.

Bianchi, L., Sami, S., Hillebrand, A., Fawcett, I. P., Quitadamo, L. R., and Seri, S. (2010). Which physiological components are more suitable for visual ERP based brain–computer interface? a preliminary MEG/EEG study. *Brain topography*, 23(2):180–185.

Choquet, G. (1953). Theory of capacities. *Annales de l'Institute Fourier*, 5:131–295.

De Campos, L. M. and Jorge, M. (1992). Characterization and comparison of Sugeno and Choquet integrals. *Fuzzy Sets and Systems*, 52(1):61–67.

Faradji, F., Ward, R. K., and Birch, G. E. (2008). Self–paced BCI using multiple SWT–based classifiers. In *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2095–2098. IEEE.

Farwell, L. A. and Donchin, E. (1988). Talking off the top of your head: toward a mental prosthesis utilizing event–related brain potentials. *Electroencephalography and Clinical Neurophysiology*, 70(6):510–523.

Grabisch, M. (1996). The application of fuzzy integrals in multicriteria decision making. *European journal of operational research*, 89(3):445–456.

Grabisch, M. (1997). *k*-order additive discrete fuzzy measures and their representation. *Fuzzy sets and systems*, 92(2):167–189.

Grabisch, M., Nguyen, H. T., and Walker, E. A. (1995). *Fundamentals of uncertainty calculi with applications to fuzzy inference*. Kluwer Academic Publishers.

Hochberg, L. R., Bacher, D., Jarosiewicz, B., Masse, N. Y., Simeral, J. D., Vogel, J., Haddadin, S., Liu, J., Cash, S. S., van der Smagt, P., and Donoghue, J. P. (2012). Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, 485(7398):372–375.

Johnson, G. D. and Krusienski, D. J. (2009). Ensemble SWLDA classifiers for the P300 speller. In Jacko,

J. A., editor, *Human–Computer Interaction. Novel Interaction Methods and Techniques*, pages 551–557. Springer.

Krusienski, D., Sellers, E., McFarland, D., Vaughan, T., and Wolpaw, J. (2008). Toward enhanced P300 speller performance. *Journal of Neuroscience Methods*, 167(1):15–21.

Kuncheva, L. (2001). Combining classifiers: Soft computing solutions". In Pal, S. and Pal, A., editors, *Pattern recognition: From classical to modern approaches*, pages 427–451. World Scientific.

Miranda, P. and Grabisch, M. (1999). Optimization issues for fuzzy measures. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 7(6):545–560.

Muller-Putz, G. R. and Pfurtscheller, G. (2008). Control of an electrical prosthesis with an SSVEP-based BCI. *IEEE Transactions on Biomedical Engineering*, 55(1):361–364.

Murofushi, T. and Soneda, S. (1993). Techniques for reading fuzzy measures (iii): Interaction index. In *9th Fuzzy Systems Symposium*, pages 693–696. In Japanese.

Murofushi, T. and Sugeno, M. (2000). Fuzzy measures and fuzzy integrals. In Grabisch, M., Murofushi, T., Sugeno, M., and Kacprzyk, J., editors, *Fuzzy Measures and Integrals - Theory and Applications*, pages 3–41. Physica Verlag.

Rakotomamonjy, A. and Guigue, V. (2008). BCI competition III: Dataset II - ensemble of SVMs for BCI P300 speller. *IEEE Transactions on Biomedical Engineering*, 55(3):1147–1154.

Ranawana, R. and Palade, V. (2006). Multi-classifier systems: Review and a roadmap for developers. *International Journal of Hybrid Intelligent Systems*, 3(1):35–61.

Rebsamen, B., Guan, C., Zhang, H., Wang, C., Teo, C., Ang, V., and Burdet, E. (2010). A brain controlled wheelchair to navigate in familiar environments. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 18(6):590–598.

Schettini, F., Aloise, F., Aric, P., Salinari, S., Mattia, D., and Cincotti, F. (2014). Self–calibration algorithm in an asynchronous P300-based brain–computer interface. *Journal of Neural Engineering*, 11(3):035004.

Sellers, E. W., Vaughan, T. M., and Wolpaw, J. R. (2010). A brain-computer interface for long-term independent home use. *Amyotrophic Lateral Sclerosis*, 11(5):449–455.

Shapley, L. (1953). A value for *n*-person games. In Kuhn, H. and Tucker, A., editors, *Contributions to the theory of games*, volume II, pages 307–317. Princeton University Press.

Sugeno, M. (1974). *Theory of fuzzy integrals and its applications*. PhD thesis, Tokyo, Japan.

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791.