Interaction-Sensitive Fuzzy Measure in Dynamic Classifier Aggregation: an Experimental Comparison

Štefka, David
2011

Interaction-Sensitive Fuzzy Measure in Dynamic Classifier Aggregation: an Experimental Comparison

Post-Graduate Student:
ING. DAVID ŠTEFKA

Department of Mathematics
Faculty of Nuclear Science and Physical Engineering
Czech Technical University
Trojanova 13
120 00 Prague 2, CZ

david.stefka@gmail.com

Supervisor:
ING. RNDR. MARTIN HOLEŇA, CSC.

Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ

martin@cs.cas.cz

Field of Study:
Mathematical Engineering

The research reported in this paper was partially supported by the grant ME949 of the Ministry of Education, Youth and Sports of the Czech Republic.

Abstract

In dynamic classifier aggregation, the fuzzy integral is used often as an aggregation operator. As the fuzzy measure of the integral, Sugeno λ-measure (which belongs to a more general class of $\perp$-decomposable fuzzy measures) is used most often. However, there is usually no explicit reason why this particular measure is used, and moreover, the measure cannot model the similarities of the individual classifiers in the team. In this paper, we show that $\perp$-decomposable measures are not appropriate for classifier combining, and we introduce the Interaction-Sensitive Fuzzy Measure (ISFM), designed specifically for classifier combining. The experiments with 3 different classifier systems on 26 benchmark datasets show that ISFM outperforms the Sugeno λ-measure in most cases.

1. Introduction

This paper is an extension of [1], in which we introduced the Interaction-Sensitive Fuzzy Measure. In this paper, we discuss the ISFM in more detail and perform more experiments.

Classifier combining methods are a popular tool for improving the quality of classification. Instead of using just one classifier, a team of classifiers is created, and the predictions of the team are combined into a single prediction [2–4]. There are two main approaches to classifier combining: *classifier selection* (where a single classifier from the team is selected for prediction according to some criterion) and *classifier aggregation* (where the outputs of the classifiers are aggregated into a single prediction). Classifier combination can be either *static*, i.e., the combining process is the same for all patterns, or *dynamic*, where the combination process is adapted to the currently classified pattern [5–9].

One of the popular aggregation operators is the *fuzzy integral* [2, 10–12]. The fuzzy integral aggregates the outputs of the individual classifiers in the team with respect to a fuzzy measure, representing the classification confidences. *Fuzzy measure* is a generalization of the additive probabilistic measure, where the additivity is replaced by a weaker condition, monotonicity – this gives us a tool which can model interactions between different elements of the fuzzy measure space. However, due to the lack of additivity, the fuzzy measure needs to be defined on all subsets of the fuzzy measure space, resulting in 2^r defining values for finite cases, where r is the size of the universe. There are several approaches to overcome this weakness: *symmetric fuzzy measures* [10], for which the value of the measure depends only on the number of elements in the argument, and $\perp$*-decomposable* fuzzy measures, including *Sugeno λ-measure* [10, 11], for which the fuzzy measure values are computed from the fuzzy measure values for the singletons (called *fuzzy densities*) using a fixed t-conorm $\perp$. However, since the value of a set of elements is computed only using the fuzzy densities of its elements and a fixed $\perp$, the similarity of the elements in the set is not taken into account, and the ability to model interactions between different elements of the fuzzy measure universe is limited.

In the literature of classifier aggregation, fuzzy integral is usually used with Sugeno λ-measure. There is usually no explicit reason for the choice of this measure other than its simplicity. Sugeno λ-measure is a special case of a $\perp$-decomposable fuzzy measure, and as such, it cannot

model similarities between the individual classifiers, and thus the contribution of using fuzzy integral is unclear.

In classifier aggregation, we usually try to create a team of classifiers that are not similar. This property is called *diversity* [13]. There are many methods for building a diverse team of classifiers [3, 14–16]; however, the team always contains classifiers that are similar. If we use the fuzzy integral with a symmetric or $\perp$-decomposable fuzzy measure, we are not able to incorporate the diversity into the measure (and thus to the aggregation process), because the fuzzy measure of a union of two sets is a function only of the fuzzy measures of the two sets, regardless of the similarity of the elements in the sets.

To overcome this weakness, we have introduced an *Interaction-Sensitive Fuzzy Measure* (ISFM) [1], which is defined using the fuzzy measure values for the singletons (fuzzy densities), and the similarities of the elements in the universe. If the fuzzy measure space corresponds to the team of classifiers, the fuzzy measure incorporates both the classification confidence (fuzzy densities), and the diversity of the team of classifiers (mutual similarities of the classifiers). Using ISFM in fuzzy integral as an aggregation operator in classifier aggregation, the aggregation process involves all the important properties: the predictions of the classifiers, the classification confidences, and the diversity of the team.

Our preliminary experiments with ISFM used with the Choquet integral in Random Forest ensembles have shown that ISFM outperforms Sugeno λ-measure [1]. In this paper, the results of a more profound investigation are reported, and the experiments have been extended to cover the Sugeno integral and also other classification models, namely ensembles of k-Nearest Neighbor classifiers [17] created by bagging [14] and ensembles of Quadratic Discriminant Classifiers [17] created by the Multiple feature subset method [18].

The paper is structured as follows. In Section 2, we briefly summarize the formalism of classification, classification confidence, and classifier combining. Section 3 describes fuzzy measures, fuzzy integrals, and their use in classifier aggregation. In Section 4, we introduce the ISFM, and in Section 5, we experimentally compare the performance of the ISFM to the performance of the Sugeno λ-measure. Section 6 then summarizes the paper.

2. Classifier Combining

In this section, we recall the formalism of dynamic classifier combining, proposed in [5]. Throughout the rest of the paper, we use the following notation. Let $\mathcal{X} \subseteq \mathbf{R}^n$ be a n-dimensional *feature space*, let $C_1, \ldots, C_N \subseteq \mathcal{X}$, $N \geq 2$ be disjoint sets called *classes*. A *pattern* is a tuple $(\vec{x}, c_{\vec{x}})$, where $\vec{x} \in \mathcal{X}$ are *features* of the pattern, and $c_{\vec{x}} \in \{1, ..., N\}$ is the index of the class the pattern belongs to. The goal of classification is to determine the class a given pattern belongs to, i.e., to predict $c_{\vec{x}}$ for unclassified patterns. We assume that for

every $\vec{x} \in \mathcal{X}$, there is a unique classification $c_{\vec{x}}$, but since it is usually not known, we will sometimes refer to a pattern only as $\vec{x} \in \mathcal{X}$.

Definition 1 *The term* classifier *denotes a mapping* $\phi : \mathcal{X} \to [0,1]^N$*, i.e., for* $\vec{x} \in \mathcal{X}$*,* $\phi(\vec{x}) = (\gamma_1(\vec{x}), \ldots, \gamma_N(\vec{x}))$*. The components* $(\gamma_1(\vec{x}), \ldots, \gamma_N(\vec{x}))$ *are called* degrees of classification (d.o.c.) *to each class.*

The d.o.c. to class C_j expresses the predicted extent to which the pattern belongs to class C_j. The prediction of $c_{\vec{x}}$ for an unknown pattern $\vec{x}$ is done by converting the continuous d.o.c. of the classifier into a *crisp output* $\phi^{(cr)}(\vec{x}) = \arg\max_{i=1,\ldots,N} \gamma_i(\vec{x})$ if there are no ties, or arbitrarily as $\phi^{(cr)}(\vec{x}) \in \arg\max_{i=1,\ldots,N} \gamma_i(\vec{x})$ in the case of ties.

2.1. Classification Confidence

In addition to the classifier output (the d.o.c.), which predicts to which class a pattern belongs, we will work with the *confidence* of the prediction, i.e., the extent to which we can "trust" the output of the classifier.

Definition 2 *Let* ϕ *be a classifer and* $\kappa_\phi : \mathcal{X} \to [0,1]$*. Then* κ_ϕ *is called a* confidence measure *and for* $\vec{x} \in \mathcal{X}$*,* $\kappa_\phi(\vec{x})$ *is called* classification confidence *of* ϕ *on* $\vec{x}$*. A confidence measure is called* static *if it is a constant of the classifier, and* dynamic *otherwise.*

The higher the trust in the classification, the closer $\kappa_\phi(\vec{x})$ is to 1. Static confidence measures evaluate the classifier as a whole and they are usually computed on a validation set after the classifier is trained. The methods include accuracy, precision, sensitivity, resemblance, etc. [17, 19]. For example, the Global Accuracy confidence measure is defined as:

$$\kappa_\phi^{(GA)} = \frac{\sum_{(\vec{y}, c_{\vec{y}}) \in \mathcal{V}} I(\phi^{(cr)}(\vec{y}) = c_{\vec{y}})}{|\mathcal{V}|}, \qquad (1)$$

where $\mathcal{V} \subseteq \mathcal{X} \times \{1, \ldots, N\}$ is the validation set and I denotes the indicator operator, defined as $I(\text{true}) = 1$, $I(\text{false}) = 0$ (we will use the notation in the rest of the paper).

Dynamic confidence measures [5–9, 20] adapt to the currently classified pattern and predict the local quality of the classification for the particular pattern $(\vec{x}, c_{\vec{x}})$. An example of a dynamic confidence measure is the Euclidean Local Accuracy (ELA):

$$\kappa_{\phi}^{(ELA)}(\vec{x}) = \frac{\sum\limits_{(\vec{y},c_{\vec{y}})\in\mathcal{V}(\vec{x})} I(\phi^{(cr)}(\vec{y}) = c_{\vec{y}})}{|\mathcal{V}(\vec{x})|}, \quad (2)$$

where $\mathcal{V}(\vec{x}) \subseteq \mathcal{V}$ is the set of validation patterns belonging to some kind of neighborhood of $\vec{x}$ (for example k nearest neighbors under Euclidean metric).

2.2. Classifier Systems

In classifier combining, instead of using just one classifier, a team of classifiers is created (sometimes called an *ensemble of classifiers*), and the team is then aggregated into one final classifier. If we want to utilize classification confidence in the aggregation process, each classifier must have its own confidence measure defined.

Definition 3 *Let $r \in \mathbf{N}$, $r \geq 2$.* Classifier team *is a tuple $(\Gamma, \mathcal{K})$, where $\Gamma = \{\phi_1, \ldots, \phi_r\}$ is a set of classifiers, and $\mathcal{K} = \{\kappa_{\phi_1}, \ldots, \kappa_{\phi_r}\}$ is a set of corresponding confidence measures.*

If a pattern $\vec{x}$ is submitted for classification, the team of classifiers returns information of two kinds – outputs of the individual classifiers (a *decision profile* [21]), and classification confidences of the classifiers on $\vec{x}$ (a *confidence vector*).

Definition 4 *Let $(\Gamma, \mathcal{K})$ be a classifier team and let $\vec{x} \in \mathcal{X}$. Then the* decision profile *of $(\Gamma, \mathcal{K})$ on $\vec{x}$ is a matrix $\Gamma(\vec{x}) \in [0,1]^{r\times N}$,*

$$\Gamma(\vec{x}) = \begin{pmatrix} \phi_1(\vec{x}) \\ \phi_2(\vec{x}) \\ \vdots \\ \phi_r(\vec{x}) \end{pmatrix} = \begin{pmatrix} \gamma_{1,1}(\vec{x}) & \gamma_{1,2}(\vec{x}) & \ldots & \gamma_{1,N}(\vec{x}) \\ \gamma_{2,1}(\vec{x}) & \gamma_{2,2}(\vec{x}) & \ldots & \gamma_{2,N}(\vec{x}) \\ & & \ddots & \\ \gamma_{r,1}(\vec{x}) & \gamma_{r,2}(\vec{x}) & \ldots & \gamma_{r,N}(\vec{x}) \end{pmatrix}, \quad (3)$$

and the confidence vector *of $(\Gamma, \mathcal{K})$ on $\vec{x}$ is a vector $\mathcal{K}(\vec{x}) \in [0,1]^r$,*

$$\mathcal{K}(\vec{x}) = \begin{pmatrix} \kappa_{\phi_1}(\vec{x}) \\ \kappa_{\phi_2}(\vec{x}) \\ \vdots \\ \kappa_{\phi_r}(\vec{x}) \end{pmatrix} \quad (4)$$

After the pattern $\vec{x}$ has been classified by all the classifiers in the team, and the confidences have been computed, these outputs have to be aggregated using a *team aggregator*. A classifier team with an aggregator will be called a *classifier system*, which can be also viewed as a single classifier.

Definition 5 *Let $(\Gamma, \mathcal{K})$ be a classifier team, and let $\mathcal{A} : [0,1]^{r\times N} \times [0,1]^r \to [0,1]^N$. The triple $\mathcal{S} = (\Gamma, \mathcal{K}, \mathcal{A})$ is called a* classifier system *and $\mathcal{A}$ is called a* team aggregator. *We define an* induced classifier *of $\mathcal{S}$ as a classifier Φ:*

$$\Phi(\vec{x}) = \mathcal{A}(\Gamma(\vec{x}), \mathcal{K}(\vec{x})) = (\gamma_1(\vec{x}), \ldots, \gamma_N(\vec{x})).$$

An example of an aggregation operator is the mean value, which defines the aggregated d.o.c. to class j as the arithmetic mean of the d.o.c. to class j given by the individual classifiers in the team:

$$\gamma_j(\vec{x}) = \frac{\sum\limits_{i=1,\ldots,r} \gamma_{i,j}(\vec{x})}{r}. \quad (5)$$

We can distinguish three types of classifier systems: *confidence-free* (which do not utilize the classification confidence at all), *static* (which use only static classification confidence), and *dynamic* (which use dynamic classification confidence, i.e., the aggregation is adapted to a particular pattern). In this paper, we are mainly interested in dynamic classifier systems.

Many aggregation operators have been studied in the literature: simple arithmetic operations (voting, sum, maximum, minimum, mean, weighted mean, weighted voting, product, etc., [21]), probability-based approaches (e.g., product rule [21], Dempster-Shafer fusion [21]), and fuzzy logic methods (fuzzy integral [12], decision templates [12, 21]). Our key interest in this paper lies in studying dynamic classifier aggregation using the fuzzy integral, which is described in the following section.

3. Fuzzy Integral, Measures and Similarity

Fuzzy integral [10, 11, 22] is an aggregation operator, based on a *fuzzy measure* (sometimes called *capacity*), which is a generalization of the additive measure, such that the additivity is replaced by a weaker condition – monotonicity. Several definitions of a fuzzy integral exists in the literature – among them, the Choquet integral and the Sugeno integral are used most often. In this section, we briefly summarize the basic definitions, and we show how the fuzzy integral can be used in classifier aggregation. For simplicity reasons, we restrict ourselves to the discrete case, and to functions in $[0, 1]$.

Definition 6 *A fuzzy measure μ on a set $\mathcal{U} = \{u_1, \ldots, u_r\}$ is a function on the power set of $\mathcal{U}$, $\mu : \mathcal{P}(\mathcal{U}) \to [0, 1]$, such that:*

1. $\mu(\emptyset) = 0, \mu(\mathcal{U}) = 1$ *(boundary conditions)*
2. $A \subseteq B \Rightarrow \mu(A) \leq \mu(B)$ *(monotonicity)*

As the universe $\mathcal{U}$ will correspond to the set of classifiers in the team, we use r to denote the universe size (cf. Sec. 3.1). We can now define the Choquet integral, which is a generalization of the classical probabilistic integral (for additive measures, it reduces to the Lebesgue integral, i.e., weighted mean in the discrete case), and the Sugeno integral. As there is no generally accepted definition of a fuzzy integral [10, 23], we restrict ourselves to the Choquet and Sugeno integrals in the rest of the paper.

We will use the following notation. Let $f : \mathcal{U} = \{u_1, \ldots, u_r\} \to [0, 1]$, $f(u_i) = f_i,\ i = 1, \ldots, r$. Then $< \cdot >$ indicates that the indices have been permuted, such that $0 = f_{<0>} \leq f_{<1>} \leq \cdots \leq f_{<r>} \leq 1$. Moreover, $A_{<i>} = \{u_{<i>}, \ldots, u_{<r>}\}$ denotes the set of of elements of $\mathcal{U}$ corresponding to the $(r - i)$ highest values of f.

Definition 7 *Let μ be a fuzzy measure on $\mathcal{U}$. Then the* Choquet integral *of a function $f : \mathcal{U} \to [0, 1]$, $f(u_i) = f_i,\ i = 1, \ldots, r$, with respect to μ is defined as:*

$$(C) \int f d\mu = \sum_{i=1}^{r} (f_{<i>} - f_{<i-1>}) \mu(A_{<i>}). \quad (6)$$

Definition 8 *Let μ be a fuzzy measure on $\mathcal{U}$. Then the* Sugeno integral *of a function $f : \mathcal{U} \to [0, 1]$, $f(u_i) = f_i,\ i = 1, \ldots, r$, with respect to μ is defined as:*

$$(S) \int f d\mu = \max_{i=1}^{r} \min(f_{<i>}, \mu(A_{<i>})). \quad (7)$$

3.1. Fuzzy Integral in Classifier Aggregation

In classifier aggregation, the universe $\mathcal{U}$ corresponds to the set of classifiers Γ in the team, i.e., $\mathcal{U} = \Gamma = \{\phi_1, \ldots, \phi_r\}$. For $\vec{x} \in \mathcal{X}$, the individual columns of the decision profile $\Gamma(\vec{x})$ are integrated using the fuzzy integral, i.e., the aggregated d.o.c. to class j is defined as

$$\gamma_j(\vec{x}) = \int \Gamma_{*,j} d\mu, \quad (8)$$

where $\int$ is a fuzzy integral, $\Gamma_{*,j}$ is the j-th column of Γ (d.o.c. to class C_j), and μ is a fuzzy measure on Γ. The fuzzy measure μ represents the importance of a particular set of classifiers used in the integration ($\mu(A_{<i>})$ represents the importance of the classifiers corresponding to the $(r - i)$ highest d.o.c.). Usually, μ somehow depends on the confidence vector $\mathcal{K}(\vec{x})$.

3.2. Important Types of Fuzzy Measures

The behavior of the fuzzy integral depends heavily on the considered fuzzy measure. As the definition of a fuzzy measure is very general, it gives us a lot of freedom when defining a fuzzy measure. However, to define a general fuzzy measure in the discrete case, we need to define all its 2^r values, which is usually very complicated. To overcome this weakness, approaches which do not need all the 2^r values have been developed [10, 11].

3.2.1 Additive Measures:

Definition 9 *Fuzzy measure μ on $\mathcal{U}$ is called* additive, *if $\mu(A \cup B) = \mu(A) + \mu(B)$ for disjoint $A, B \subseteq \mathcal{U}$.*

Additive measures correspond to the classical probabilistic measures. The measure is defined only using the values for the singletons, $\mu(\{u_i\}),\ i = 1, \ldots, r$ (called *fuzzy densities*), and all the remaining values are computed using the additivity condition. However, such measure cannot model interaction between the elements of the fuzzy measure space (which in particular implies that the diversity of the team of classifiers cannot be taken into account in the aggregation). Choquet integral with an additive measure reduces to the weighted mean.

3.2.2 Symmetric Measures:

Definition 10 *Fuzzy measure μ on $\mathcal{U}$ is called* symmetric, *if for $A, B \subseteq \mathcal{U}$, $|A| = |B| \Rightarrow \mu(A) = \mu(B)$, i.e., its value depends only on the cardinality of the argument, $\mu(A) = g(|A|)$.*

We can choose any nondecreasing function g, such that $g(0) = 0$ and $g(r) = 1$ to model the importance of a set of r elements. If a symmetric measure is used in Choquet integral, the integral reduces to the Ordered Weighted Average operator [10]. However, symmetric measures assume that all the classifiers have the same importance, and thus not only symmetric fuzzy measures do not take similarities of the classifiers into account, but moreover, the resulting aggregation scheme is confidence-free, i.e., the classificatoin confidence does not influence the aggregation. As we deal with dynamic classifier systems only, we do not take symmetric measures into account in the rest of the paper.

3.2.3 $\perp$-decomposable Measures:

Definition 11 *Let μ be a fuzzy measure on $\mathcal{U}$ and let $\perp$ be a t-conorm. Then μ is called* $\perp$-decomposable*, if for disjoint $A, B \subseteq \mathcal{U}$,*

$$\mu(A \cup B) = \mu(A) \perp \mu(B). \tag{9}$$

$\perp$-decomposable measures need only the r fuzzy densities and all the other values are computed using the formula (9). Particular cases of $\perp$-decomposable fuzzy measures are additive measures ($\perp$ being the bounded sum), and the Sugeno λ-measure [10, 11], defined as

$$\mu(A \cup B) = \mu(A) + \mu(B) + \lambda\mu(A)\mu(B), \tag{10}$$

for disjoint $A, B \in \mathcal{U}$, and some fixed $\lambda > -1$. The value of λ is computed as the unique non-zero root greater than -1 of the equation

$$\lambda + 1 = \prod_{i=1,\ldots,r} (1 + \lambda\mu(\{u_i\})), \tag{11}$$

if the densities do not sum to 1. If they do sum to 1, $\lambda = 0$ and the fuzzy measure is additive.

The Sugeno λ-measure is used most often in classifier aggregation using fuzzy integral (with the fuzzy densities corresponding to the classification confidences, $\mu(\{u_i\}) = \kappa_{\phi_i}(\vec{x})$). However, its use is usually not supported by any arguments and it is basically selected because of its simplicity.

A strong weakness of any $\perp$-decomposable measure (and Sugeno λ-measure in particular) is that it cannot model the interaction (similarities) between the classifiers, because the fuzzy measure value of a set of two (or more) classifiers is fully determined by the formula (9) with a fixed $\perp$. Therefore, the diversity of the team of classifiers cannot be taken into account in the aggregation (as in the case of additive measures).

To overcome the weaknesses of the methods presented above, we have defined an Interaction-Sensitive Fuzzy Measure (ISFM) [1], which is defined not only using the fuzzy densities, but also using mutual similarities of the classifiers in the team. The method is described in the following section, but prior to that, we formally define the concept of a similarity [24].

3.3. Similarity of Classifiers

Definition 12 *Let $\wedge$ be a t-norm and let $S : \mathcal{U} \times \mathcal{U} \to [0, 1]$ be a fuzzy relation. S is called a* similarity with respect to $\wedge$ *if the following holds $\forall a, b, c \in \mathcal{U}$:*

- $S(a, a) = 1$ *(reflexivity)*
- $S(a, b) = S(b, a)$ *(symmetry)*
- $S(a, b) \wedge S(b, c) \leq S(a, c)$ *(transitivity w.r.t. $\wedge$)*

In the context of classifier combining, we will work with similarity of classifiers in particular, which, for classifiers ϕ_k, ϕ_l, will be measured empirically as the proportion of equal crisp predictions on the validation set $\mathcal{V}$,

$$S(\phi_k, \phi_l) = \frac{\sum_{(\vec{y}, c_{\vec{y}}) \in \mathcal{V}} I(\phi_k^{(cr)}(\vec{y}) = \phi_l^{(cr)}(\vec{y}))}{|\mathcal{V}|}. \tag{12}$$

The relation (12) is a similarity with respect to Łukasiewicz t-norm $\wedge_L$, but it is not a similarity with respect to standard or product t-norms $\wedge_S$, $\wedge_P$.

4. Interaction-Sensitive Fuzzy Measure and its Use in Fuzzy Integral

Methods for constructing a team of classifiers usually try to create a team which is both both *accurate* and *diverse* [2, 3, 13]. Diversity of the classifiers in the team is a key property in classifier combining, since if the classifiers are very similar, the classifier combining cannot improve the classification quality.

Fuzzy measures represent a convenient tool to work with the diversity of the team. As $\mu(A_{<i>})$ are computed for $i = r, \ldots, 1$, i.e., in i-th step, classifier $\phi_{<i>}$ is added to the set of classifiers $A_{<i+1>} = \{\phi_{<i+1>}, \ldots, \phi_{<r>}\}$, we can influence the increase of the fuzzy measure – if $\phi_{<i>}$ is similar to the classifiers in $A_{<i+1>}$, the increase in the fuzzy measure should be small (since the importance of the set $A_{<i>}$ should be similar to the importance of the set $A_{<i+1>}$), and if $\phi_{<i>}$ is not similar to the classifiers in $A_{<i+1>}$, the increase of the fuzzy measure should be large.

$\perp$-decomposable fuzzy measures (and in particular additive measures and Sugeno λ-measure) cannot model such interactions between the classifiers, because they are defined only using the fuzzy densities and a fixed $\perp$. Therefore, we propose an *Interaction-Sensitive Fuzzy Measure* (ISFM), which incorporates the similarities of the classifiers in the team, defined using the following recursive definition.

Definition 13 *Let $\mathcal{U} = \{u_1, \ldots, u_r\}$ be a universe, let S be a similarity w.r.t. a t-norm $\wedge$, $s_{i,j} = S(u_i, u_j)$, let $\kappa_i \in [0, 1]$, $i = 1, \ldots, r$ denote the importance (weight) of u_i, and let $A_{<i>} = \{u_{<i>}, \ldots, u_{<r>}\}$,*

$A_{<r+1>} = \emptyset$, where $< \cdot >$ denotes index ordering according to some $f : \mathcal{U} \to [0,1]$, such that $0 \leq f_{<1>} \leq \cdots \leq f_{<r>} \leq 1$.

Let $\widetilde{\mu} : \mathcal{P}(\mathcal{U}) \to \mathbf{R}^+$, such that

$$\widetilde{\mu}(\emptyset) = 0 \tag{13}$$

$$\widetilde{\mu}(A_{<r>}) = \widetilde{\mu}(\{u_{<r>}\}) = \kappa_{<r>} \tag{14}$$

$$\widetilde{\mu}(A_{<i>}) = \widetilde{\mu}(\{u_{<i>}, \ldots, u_{<r>}\}) = \tag{15}$$

$$= \widetilde{\mu}(A_{<i+1>}) + (1 - \max_{k=i+1}^{r} s_{<i>,<k>})\kappa_{<i>} \tag{16}$$

$$\text{for } i = r-1, \ldots, 1, \tag{17}$$

and $\forall X \subseteq \mathcal{U},\ X \neq A_{<i>},\ i = 1, \ldots, r,$

$$\widetilde{\mu}(X) = \widetilde{\mu}(A_{<q>}), \tag{18}$$

where $q = \min\{i = r+1, \ldots, 1 | A_{<i>} \subseteq X\}$.

The mapping $\mu^{(I)} : \mathcal{P}(\mathcal{U}) \to [0,1]$, defined as

$$\mu^{(I)}(X) = \frac{\widetilde{\mu}(X)}{\widetilde{\mu}(A_{<1>})} = \frac{\widetilde{\mu}(X)}{\widetilde{\mu}(\mathcal{U})}, \tag{19}$$

is called an Interaction-Sensitive Fuzzy Measure (ISFM) *on $\mathcal{U}$ with respect to f.*

For the fuzzy integration itself, only the values for $A_{<i>}$, $i = 1, \ldots, r$ (15-17) are needed, the remaining values (18) represent an extension to the whole power set and are needed only for $\mu^{(I)}$ to be properly defined. (19) represents a normalization of $\widetilde{\mu}$ to [0,1].

The definition is general and can be used also in other applications than classifier combining. In classifier combining, $\mathcal{U} = \Gamma$ is the set of classifiers, $\kappa_i = \kappa_\phi(\vec{x})$ are the classification confidences, $f = \Gamma_{*,j}$ is the j-th column of the decision profile, and S denotes the similarity of classifiers (12). The following proposition shows that $\mu^{(I)}$ is well-defined.

Proposition 1 *$\mu^{(I)}$ is a fuzzy measure on $\mathcal{U}$.*

Proof: The boundary conditions follow directly from the definition of $\mu^{(I)}$. Let $X \subseteq Y \subseteq \mathcal{U}$. Then $q_X = \min\{i = r+1, \ldots, 1 | A_{<i>} \subseteq X\} \geq q_Y = \min\{i = r+1, \ldots, 1 | A_{<i>} \subseteq Y\}$, and thus, $\mu^{(I)}(X) = \mu^{(I)}(A_{<q_X>}) \leq \mu^{(I)}(A_{<q_Y>}) = \mu^{(I)}(Y)$, which proves the monotonicity. ∎

In (16), $\max_{k=i+1}^{r} s_{<i>,<k>}$ incorporates the diversity of the team of classifiers into the fuzzy measure. The following proposition shows that if for some i, the i-th classifier is totally similar to some other classifier in $A_{<i+1>}$, then $\mu^{(I)}$ does not increase, and if it is totally unsimilar to all classifiers in $A_{<i+1>}$, the increase in the fuzzy measure is maximal.

Proposition 2 *Let $\mu^{(I)}$ be an ISFM on $\mathcal{U}$ w.r.t. $f : \mathcal{U} \to [0,1]$, and let $i \in \{1, \ldots, r-1\}$. Then the following holds*

1. $\exists k \in \{i+1, \ldots, r\}\ s_{<i>,<k>} = 1 \Rightarrow \mu^{(I)}(A_{<i>}) = \mu^{(I)}(A_{<i+1>})$
2. $\forall k \in \{i+1, \ldots, r\}\ s_{<i>,<k>} = 0 \Rightarrow \mu^{(I)}(A_{<i>}) = \mu^{(I)}(A_{<i+1>}) + \kappa_{<i>}/\widetilde{\mu}(\mathcal{U})$

Proof: Trivially from (16) and (19). ∎

The following proposition describes an extreme case, in which all the classifiers are totally similar (the measure in the integral behaves like a constant measure and Choquet and Sugeno intergrals reduce to the maximum value).

Proposition 3 *Let $\mu^{(I)}$ be an ISFM on $\mathcal{U}$ w.r.t. $f : \mathcal{U} \to [0,1]$, $f(u_i) = f_i$, and let $\forall i, j \in \{1, \ldots, r\},\ i \neq j,\ s_{i,j} = 1$. Then $\forall X \subseteq \mathcal{U}$*

1. $\forall k \in \{1, \ldots, r\}\ \mu^{(I)}(A_{<k>}) = 1$
2. $\exists k \in \{1, \ldots, r\}\ A_{<k>} \subseteq X \Rightarrow \mu^{(I)}(X) = 1$
3. $\forall k \in \{1, \ldots, r\}\ A_{<k>} \nsubseteq X \Rightarrow \mu^{(I)}(X) = 0$
4. $(C)\int f d\mu^{(I)} = (S)\int f d\mu^{(I)} = \max_{i=1}^{r} f_i$

Proof: (1) follows directly from (15-17) and (19); (2), (3) from (18) and (4) is an application of the measure to the definition of Choquet and Sugeno integrals. ∎

Another extreme case is that all the classifiers are totally unsimilar (the measure in the integral behaves like an additive measure and the Choquet integral reduces to the weighted mean).

Proposition 4 *Let $\mu^{(I)}$ be an ISFM on $\mathcal{U}$ w.r.t. $f : \mathcal{U} \to [0,1]$, $f(u_i) = f_i$, and let $\forall i, j \in \{1, \ldots, r\},\ i \neq j,\ s_{i,j} = 0$. Then the following holds:*

1. $\forall k \in \{1, \ldots, r\}\ \mu^{(I)}(A_{<k>}) = \frac{\sum_{l=k}^{r} \kappa_{<l>}}{\widetilde{\mu}(\mathcal{U})} = \frac{\sum_{l=k}^{r} \kappa_{<l>}}{\sum_{l=1}^{r} \kappa_{<l>}}$

2. $(C)\int f d\mu^{(I)} = \frac{\sum_{l=1}^{r} f_{<l>}\kappa_{<l>}}{\sum_{l=1}^{r}\kappa_{<l>}}$

Proof: (1) follows directly from (15-17) and (19). (2) is an application of the measure to the definition of the Choquet integral. ■

5. Experiments

To experimentally compare the ISFM-based approach with the Sugeno λ-measure approach, we designed three different classifier systems:

- Random Forest ensemble [16]. In our experiments, we used $r = 20$ trees.
- Ensemble of k-Nearest neighbor classifiers [17] created by bagging [14]. In our experiments, we used $r = 20$ classifiers in the team with $k = 5$.
- Ensemble of Quadratic discriminant classifiers [17] created by the multiple feature subset method [18]. Each classifier was trained only on a subset of features. For datasets with $n \leq 5$ dimensions, all possible subsets (feature combinations) in the MFS were used. For higher dimensional datasets, 32 subsets of features were selected by bagging.

To compute the classification confidence, we used the ELA method (2). The number of neighbors was set based on the size of the dataset to $k = 5$ (≤ 500 patterns), $k = 10$ ($501 - 1000$ patterns), or $k = 20$ (> 1000 patterns). The values of the parameters were set based on preliminary testing, no optimization or fine-tuning was done. As aggregation operators, we used the following

- Weighted mean – representing the baseline (special case of the Choquet integral with additive measure)
- Choquet integral with the λ-measure
- Choquet integral with the ISFM
- Sugeno integral with the λ-measure
- Sugeno integral with the ISFM
- Single best (for reference) – mean error rate of the classifier with lowest error rate selected in each crossvalidation run, representing the "worst-case" scenario
- Oracle (for reference) – the theoretical "best-case" scenario, which, for a given pattern, gives correct prediction if and only if any of the classifiers in the team gives correct prediction

The methods were implemented in the Java programming language and the experiment was performed on 7 artificial and 19 real-world datasets with varying size, dimensionality, and class count (due to numerical instabilities of the QDC model, we had to leave out three real-world datasets for the QDC ensemble). The properties of the datasets are shown in Table 1. We used 10-fold cross-validation to measure the performance of the methods (8 folds for training set, 9th fold for validation set, 10th fold for testing set, with cyclic shift). The validation set was used to compute the classification confidence and the similarity of the classifiers in the team, and the testing set was used to compare the results of the methods. The mean value and standard deviation of the error rate were measured. We also measured statistical significance of the results (at 5% confidence level by the analysis of variance using Tukey-Kramer method).

Table 1: Properties of the datasets used in the experiments.

Dataset	ref.	size	classes	dimensions
Artificial				
clouds	[25]	5000	2	2
concentric	[25]	2500	2	2
gauss 3D	[25]	5000	2	3
gauss 8D	[25]	5000	2	8
ringnorm	[26]	3000	2	20
twonorm	[26]	3000	2	20
waveform	[26]	5000	3	21
Real-world				
balance	[26]	625	3	4
breast	[26]	699	2	9
glass	[26]	214	7	9
iris	[26]	150	3	4
letter-recg.	[26]	20000	26	16
pendigits	[26]	10992	10	16
phoneme	[25]	5427	2	5
pima	[26]	768	2	8
poker	[26]	4828	3	10
satimage	[25]	6435	6	4
segmentation	[26]	2310	7	16
sonar	[26]	208	2	60
texture	[25]	5500	11	10
transfusion	[26]	748	2	4
vehicle	[26]	946	4	18
vowel	[26]	990	11	10
wine	[26]	178	3	13
wineq-red	[26]	1600	3	11
wineq-white	[26]	4898	3	11
yeast	[26]	1484	4	8

Table 2: Random Forest: The i, j-th element of the table shows the number of datasets in which method i obtained lower mean error rate than method j. The number in parentheses, if present, shows the number of datasets for which the improvement was statistically significant (excluding Oracle). The last column shows the number of datasets for which a given method was better than all the other methods (excluding Oracle).

↓ superior to →	SB	WMean	CI-λ	CI-ISFM	SI-λ	SI-ISFM	Oracle	all
SB	-	0	1 (1)	0	1 (1)	0	0	0
WMean	26 (16)	-	12 (3)	3	12 (5)	5	0	1
CI-λ	25 (16)	14	-	5	14	8	0	1
CI-ISFM	26 (18)	23	21 (5)	-	19 (5)	16	0	11
SI-λ	25 (17)	14	12	6	-	8	0	4
SI-ISFM	26 (18)	21	18 (3)	10	18 (4)	-	0	9
Oracle	26	26	26	26	26	26	-	26

Table 3: k-NN ensemble: The i, j-th element of the table shows the number of datasets in which method i obtained lower mean error rate than method j. The number in parentheses, if present, shows the number of datasets for which the improvement was statistically significant (excluding Oracle). The last column shows the number of datasets for which a given method was better than all the other methods (excluding Oracle).

↓ superior to →	SB	WMean	CI-λ	CI-ISFM	SI-λ	SI-ISFM	Oracle	all
SB	-	7	3	2	2	2	0	0
WMean	19 (1)	-	3	4	3	3	0	0
CI-λ	23 (3)	23	-	10	17	11	0	9
CI-ISFM	24 (6)	22 (3)	16	-	19 (1)	14	0	10
SI-λ	25 (2)	23 (1)	11	7	-	8	0	2
SI-ISFM	24 (8)	23 (3)	15	12	18 (1)	-	0	7
Oracle	26	26	26	26	26	26	-	26

Table 4: QDC ensemble: The i, j-th element of the table shows the number of datasets in which method i obtained lower mean error rate than method j. The number in parentheses, if present, shows the number of datasets for which the improvement was statistically significant (excluding Oracle). The last column shows the number of datasets for which a given method was better than all the other methods (excluding Oracle).

↓ superior to →	SB	WMean	CI-λ	CI-ISFM	SI-λ	SI-ISFM	Oracle	all
SB	-	8	6	4	7	3	0	1
WMean	15 (8)	-	12 (2)	2	13 (2)	4	0	0
CI-λ	17 (6)	11	-	5	14 (1)	7	0	3
CI-ISFM	19 (8)	21 (4)	19 (5)	-	19 (5)	11	0	10
SI-λ	16 (8)	10	9	4	-	7	0	1
SI-ISFM	20 (9)	19 (4)	16 (5)	12	16 (5)	-	0	8
Oracle	23	23	23	23	23	23	-	23

To compare the methods in general, we measured the number of datasets in which a given method outperformed other methods, the results are shown in Tables 2–4.

As our main goal in this experiment was to compare ISFM with Sugeno λ-measure, we can say the following. For the Random Forests with Choquet integral, ISFM outperformed λ-measure on 21 datasets (5 times significant), with Sugeno integral on 18 datasets (4 times significant), out of 26 datasets total. For the k-NN ensemble with Choquet integral, ISFM outperformed λ measure on 16 datasets (none significant), with Sugeno integral on 18 datasets (once significant), out of 26 datasets total. For the QDC ensemble with Choquet integral, ISFM outperformed λ measure on 19 datasets (5 times significant), with Sugeno integral on 16 datasets (6 times significant), out of 23 datasets total.

Generally speaking, fuzzy integral with ISFM usually outperformed λ-measure in most cases (sometimes statistically significantly, but no significant outperforming of λ-measure over ISFM occurred). The Choquet integral obtained slightly better results than the Sugeno integral, and the Choquet integral with ISFM was the most succesfull aggregation scheme in these experiments. Another interesting result is that while both Choquet and Sugeno integrals with ISFM outperformed the Weighted Mean, this is not true for the case of Sugeno λ-measure – in most cases, both Choquet and Sugeno integrals with λ-measure obtained comparable or significantly worse results than the Weighted mean.

6. Conclusion

In this paper, we have summarized how the fuzzy integral can be used as an aggregation operator in dynamic classifier systems. We have discussed that symmetric, and $\perp$-decomposable fuzzy measures are not appropriate for using in classifier combining with fuzzy integral and we have introduced an interaction-sensitive fuzzy measure (ISFM), which tries to overcome the weaknesses of these methods. IFSM, designed specifically for the use in classifier aggregation, provides a convenient tool for representing the diversity of the team of classifiers, and, when used in the fuzzy integral, the aggregation can incorporate the classifier predictions, the classification confidences, and also the diversity of the team. Our experiments with three different dynamic classifier systems with the Choquet and Sugeno integrals on 26 datasets show that the ISFM outperforms the Sugeno λ-measure, which is used most often in the literature in connection with the fuzzy integral.

References

[1] D. Štefka and M. Holeňa, "Dynamic classifier aggregation using fuzzy integral with interaction-sensitive fuzzy measure," in *Proceedings of the 10th International Conference on Intelligent Systems Design and Applications, ISDA 2010, November 29 - December 1, 2010, Cairo, Egypt*, pp. 225-230, IEEE, 2010.

[2] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[3] L. Rokach, "Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography," *Comput. Stat. Data Anal.*, Vol. 53, No. 12, pp. 4046-4072, 2009.

[4] D. Ruta and B. Gabrys, "An overview of classifier fusion methods," *Computing and Information Systems*, Vol. 7, pp. 1-10, 2000.

[5] D. Štefka and M. Holeňa, "Dynamic classifier systems and their applications to random forest ensembles," in *Proceedings of the ICANNGA 2009 Ninth International Conference on Adaptive and Natural Computing Algorithms, Kuopio, Finland*, vol. 5495 of *Lecture Notes in Computer Science*, p. 458-468, Springer, 2009.

[6] G. Giacinto and F. Roli, "Dynamic classifier selection based on multiple classifier behaviour," *Pattern Recognition*, Vol. 34, No. 9, pp. 1879-1881, 2001.

[7] A. H. R. Ko, R. Sabourin, and A. S. Britto, Jr., "From dynamic classifier selection to dynamic ensemble selection," *Pattern Recogn.*, Vol. 41, No. 5, pp. 1718-1731, 2008.

[8] K. Woods, J. W. Philip Kegelmeyer, and K. Bowyer, "Combination of multiple classifiers using local accuracy estimates," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 19, No. 4, pp. 405-410, 1997.

[9] A. Tsymbal, M. Pechenizkiy, and P. Cunningham, "Dynamic integration with random forests.," in *ECML* (J. Fürnkranz, T. Scheffer, and M. Spiliopoulou, eds.), Vol. 4212 of *Lecture Notes in Computer Science*, pp. 801-808, Springer, 2006.

[10] V. Torra and Y. Narukawa, *Modeling Decisions: Information Fusion and Aggregation Operators*. Springer, 2007.

[11] M. Grabisch and H. T. Nguyen, *Fundamentals of Uncertainty Calculi with Applications to Fuzzy Inference*. Norwell, MA, USA: Kluwer Academic Publishers, 1994.

[12] L. I. Kuncheva, "Fuzzy versus nonfuzzy in combining classifiers designed by boosting," *IEEE Transactions on Fuzzy Systems*, Vol. 11, No. 6, pp. 729-741, 2003.

[13] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles," *Machine Learning*, Vol. 51, pp. 181-207, 2003.

[14] L. Breiman, "Bagging predictors," *Machine Learning*, Vol. 24, No. 2, pp. 123-140, 1996.

[15] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *International Conference on Machine Learning*, pp. 148-156, 1996.

[16] L. Breiman, "Random forests," *Machine Learning*, Vol. 45, No. 1, pp. 5-32, 2001.

[17] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[18] S. D. Bay, "Nearest neighbor classification from multiple feature subsets," *Intelligent Data Analysis*, Vol. 3, No. 3, pp. 191-209, 1999.

[19] D. J. Hand, *Construction and Assessment of Classification Rules*. Wiley, 1997.

[20] S. J. Delany, P. Cunningham, D. Doyle, and A. Zamolotskikh, "Generating estimates of classification confidence for a case-based spam filter," in *Case-Based Reasoning, Research and Development, 6th Int. Conf., ICCBR 2005, Chicago, USA* (H. Muñoz-Avila and F. Ricci, eds.), Vol. 3620 of *LNCS*, pp. 177-190, Springer, 2005.

[21] L. I. Kuncheva, J. C. Bezdek, and R. P. W. Duin, "Decision templates for multiple classifier fusion: an experimental comparison.," *Pattern Recognition*, Vol. 34, No. 2, pp. 299-314, 2001.

[22] T. Murofushi and M. Sugeno, "Fuzzy t-conorm integral with respect to fuzzy measures: Generalization of Sugeno integral and Choquet integral," *Fuzzy Sets and Systems*, Vol. 42, No. 1, pp. 57-71, 1991.

[23] E. P. Klement, R. Mesiar, and E. Pap, "A universal integral as common frame for choquet and sugeno integral," *Trans. Fuz Sys.*, Vol. 18, pp. 178-187, 2010.

[24] P. Hájek, *Metamathematics of Fuzzy Logic*. Trends in Logic, Kluwer, 2001.

[25] UCL MLG, "Elena database," 1995. http://www.dice.ucl.ac.be/mlg/?page=Elena.

[26] C. B. D.J. Newman, S. Hettich and C. Merz, "UCI repository of machine learning databases," 1998. http://www.ics.uci.edu/ mlearn/MLRepository.html.