Combining Support Vector Machines by Means of Fuzzy Aggregation

MARTIN HOLEŇA, JAROSLAV MORAVEC
Institute of Computer Science, Academy of Sciences of the Czech Republic,
Pod vodárenskou věží 2, 18207 Praha 8
CZECH REPUBLIC

Abstract: The paper deals with a recently proposed approach to combining classifiers by means of fuzzy aggregation. The approach relies on the quasi-Sugeno integral and on the t-conorm integral as a generalization of the Choquet and Sugeno integral, which have been used for combining classifiers so far. New theoretical development is presented, in particular a proposition concerning the λ measures used in the quasi-Sugeno integral, and the approach is elaborated specifically for support vector machines. Finally, experience is reported that was gained when using the approach to combine support vector machines in a neurophysiologic application.

Key–Words: Support vector machines, Fuzzy aggregation, t-conorm integral, Quasi-Sugeno integral

1 Introduction

Classification, i.e., assigning objects to predefined classes based on data about those objects, is a ubiquitous task, encountered in nearly any application area. Traditionally, dealing with data was the domain of *statistics*, therefore the earliest classification methods originated there, e.g., the Bayesian classifier, linear and quadratic discriminant analysis, or nearest neighbour methods [8, 9]. Since the advent of computers, the classification task has also been studied in connection with the emerging paradigm of *machine learning*. Most successful among methods developed in that area have been various kinds of classification trees and neural networks [1, 9, 14]. Since the 1990s, the spectrum of classification methods has been complemented with *support vector machines (SVMs)* , which attempt to combine machine learning with the statistical approach [2, 15].

Besides looking for new kinds of classifiers, the efforts to improve classification followed also the way of *combining* the results of a whole set of classifiers. Particular attention has been paid to combining sets of classifiers of the same type, differing only in some tunable parameter and / or the data used for their training, which are called *ensembles of classifiers*. To this end, various approaches can be employed, from simple arithmetic rules, through using second level classifiers and boosting, to sophisticated methods such as decision templates, Dempster-Shafer fusion, or fuzzy integral [10, 12]. From the various kinds of fuzzy integral, *Choquet integral* and *Sugeno integral* have been used for combining classifiers so far.

Both the Choquet and the Sugeno integral are only particular cases of t*-conorm integrals*, the latter is at the same time a particular case of *quasi-Sugeno integrals* [7]. This fact incited research into an approach to combining classifiers based on quasi-Sugeno integrals. First results of that research have been reported in [17]. In the present paper, the approach is further developed, and elaborated specifically for support vector machines.

The following section recalls the t-conorm integrals as a general means for fuzzy aggregation and quasi-Sugeno integrals as their particular class. The key section of the paper is Section 3, in which the proposed approach to combining SVMs is described. Finally in Section 4, experience from a neurophysiologic application is reported.

2 Fuzzy Aggregation Based on the t-conorm Integral

Like traditional integrals, also fuzzy integrals aggregate all function values of a function into one single characteristics. However, they aggregate the individual function values using a more general kind of measures – fuzzy measures. The first integral of that kind was proposed already more than a decade before the advent of the fuzzy set theory by G. Choquet [3]. Within the framework of the fuzzy set theory, a number of fuzzy integrals have been elaborated, most frequently used being the one proposed 1974 by M. Sugeno [16]. In the late 1980s, T. Murofushi introduced a very broad class of t-conorm integrals, which covered almost all existing fuzzy integrals and has several important subclasses, such as Archimedean t-conorm integrals or quasi-Sugeno integrals, a generalization of Sugeno integrals [6, 7].

The present section recalls the basic concepts pertaining to the t-conorm and quasi-Sugeno integrals, as well as their relationship to the Choquet and Sugeno integrals. Since the concern of this paper is combining classifiers, only integration over finite sets is treated here.

Definition 1 *Let $\lambda > -1$, Ω be a nonempty set, $\mathcal{A}$ a set of subsets of Ω, and $\mu : \mathcal{A} \to [0,\infty)$ a mapping such that:*
(i) $\emptyset, \Omega \in \mathcal{A}$*;*
(ii) $\mu(\emptyset) = 0$*;*
(iii) $(\forall A, B \in \mathcal{A})\ A \subset B \Rightarrow \mu(A) \leq \mu(A)$.
Then

a) *μ is called a* fuzzy measure *on $(\Omega, \mathcal{A})$;*

b) *if $\mathcal{A}$ is the powerset $\mathcal{P}(\Omega)$ of Ω and μ fulfils*

$$(\forall A, B)\ A \cap B = \emptyset \Rightarrow \mu(A \cup B) = \\ = \mu(A) + \mu(B) + \lambda\mu(A)\mu(B), \quad (1)$$

it is called λ measure on Ω.

Observe that a λ measure with $\lambda = 0$ and $\mu(\Omega) = 1$ is a probability on $(\Omega, \mathcal{P}(\Omega))$.

Definition 2 *Let $\triangle$, $\perp$ and $\underline{\perp}$ be continuous t-conorms such that each of them is either Archimedean or the standard conorm $\vee$ (i.e., $x \vee y = \max(x, y)$), and $\diamond : [0,1]^2 \to [0,1]$ a binary operation such that:*
(i) *$\diamond$ is nondecreasing in both arguments;*
(ii) *$\diamond$ is continuous on $(0,1]^2$;*
(iii) $(\forall a, x \in [0,1])\ a \diamond x = 0$ *iff* $a = 0$ *or* $x = 0$*;*
(iv) $(\forall a, x, y \in [0,1])\ x \perp y < 1 \Rightarrow a \diamond (x \perp y) = (a \diamond x) \underline{\perp} (a \diamond y)$*;*
(v) $(\forall a, b, x \in [0,1])\ a \triangle b < 1 \Rightarrow (a \triangle b) \diamond x = (a \diamond x) \underline{\perp} (b \diamond x)$.
Then $\mathcal{F} = (\triangle, \perp, \underline{\perp}, \diamond)$ is called t-conorm system for integration *with* integrand space $([0,1], \triangle)$, measure space $([0,1], \perp)$, integral space $([0,1], \underline{\perp})$, *and* quasiproduct $\diamond$. *If all $\triangle$, $\perp$ and $\underline{\perp}$ are Archimedean, $\mathcal{F}$ is called* Archimedean.

Definition 3 *Let μ be a fuzzy measure on $(\Omega, \mathcal{A})$ with Ω finite and $\mu(\Omega) = 1$, and $\mathcal{F} = (\triangle, \perp, \underline{\perp}, \diamond)$ be a t-conorm system for integration. Let further $f : \Omega \to [0,1]$, $a_1, \ldots, a_n \in [0,1]$ be such that $a_1 \leq \cdots \leq a_n$ and $f(\Omega) = \{a_1, \ldots, a_n\}$. For $i = 1, \ldots, n$, denote $D_i = f^{-1}(a_i)$, $A_i = \bigcup_{j=i}^{n} D_j$. Then*

a) *the* fuzzy t-conorm integral *of f based on $\mathcal{F}$ with respect to μ is defined by*

$$(\mathcal{F}) \int f \diamond d\mu = \underline{\perp}_{i=1}^{n} (a_i -_{\triangle} a_{i-1}) \diamond \mu(A_i), \quad (2)$$

where $a_0 = 0$ and $-_{\triangle}$ denotes the pseudo-difference with respect to $\triangle$, i.e.,

$$(\forall a, b \in [0,1])\ a -_{\triangle} b = \inf\{c : b \triangle c \geq a\};$$

b) *if in particular, $\mathcal{F} = (\vee, \vee, \vee, \diamond)$, where $\diamond$ in addition to satisfying the conditions (i)–(v) from Definition 2 is a t-norm, $(\mathcal{F}) \int f \diamond d\mu$ is called* quasi-Sugeno integral.

Using the definitions of Sugeno and Choquet integral (cf. [7]), the following proposition is easy to prove:

Proposition 4 *Let μ, $\mathcal{F}$, f, and a_i, A_i for $i = 1, \ldots, n$ have the same meaning as in Definition 3. Denote $\wedge$ the standard norm (i.e., $x \wedge y = \min(x, y)$), and $\vee_{Ł}$ the Łukasiewicz conorm (i.e., $x \vee_{Ł} y = \min(1, x + y)$). Then*

1. *if in particular, $\mathcal{F} = (\vee, \vee, \vee, \wedge)$,*

$$(\mathcal{F}) \int f \diamond d\mu = \bigvee_{i=1}^{n} (a_i \wedge \mu(A_i)), \quad (3)$$

which is the definition of $\int_S f d\mu$, the Sugeno integral *of f with respect to μ;*

2. *if in particular, $\mathcal{F} = (\vee_{Ł}, \vee_{Ł}, \vee_{Ł}, \cdot)$ with $\cdot$ denoting the ordinary multiplication,*

$$(\mathcal{F}) \int f \diamond d\mu = \sum_{i=1}^{n} (a_i - a_{i-1}) \cdot \mu(A_i), \quad (4)$$

which is one of equivalent definitions of $\int_C f d\mu$, the Choquet integral *of f with respect to μ.*

3 Using Quasi-Sugeno Integral to Combine SVMs

To elaborate the use of t-conorm integral and quasi-Sugeno integral for combining support vector machines entails, according to Definition 3, the following steps:

1. to determine the sets Ω and $\mathcal{A}$;
2. to define a fuzzy measure on $(\Omega, \mathcal{A})$;
3. to choose a t-conorm system on which the integral should be based;

4. to define the integrand.

In the present section, those four steps will be individually dealt with. The first and third of them actually do not depend on the fact that specifically SVMs have been chosen as classifiers. Therefore, it is possible to make use of recently published results [17], which concern combing classifiers by means of t-conorm integral in general.

1. Sets Ω and $\mathcal{A}$. Because the objective of the integration is combining support vector machines, a natural choice for Ω is the set of considered SVMs: $\Omega = \{\varsigma_1, \ldots, \varsigma_n\}$, where $\varsigma_1, \ldots, \varsigma_n : \mathcal{X} \to \{1, -1\}$, and $\mathcal{X}$ is some Euclidian space. For finite Ω, it is common to define the domain of a fuzzy measure as $\mathcal{P}(\Omega)$ since due to the finiteness of Ω, there is no problem in obtaining any of the subsets of Ω. Moreover, the domain $\mathcal{P}(\Omega)$ is required if a λ measure is considered, according to Definition 1. Hence, our choice is $\mathcal{A} = \mathcal{P}(\Omega)$.

2. Fuzzy measure. To define a general fuzzy measure μ, values $\mu(A)$ for all the $2^n - 1$ nonempty sets $A \subset \Omega$ are needed. On the other hand if μ is restricted to be a λ measure, then the n values $\mu(\varsigma_1), \ldots, \mu(\varsigma_n)$ are sufficient, as the following proposition shows.

Proposition 5 *Let $\Omega = \{\varsigma_1, \ldots, \varsigma_n\}$. Let further $c_1, \ldots, c_n \in [0, 1]$ be such that $c_1 + \cdots + c_n > 0$. Then:*

a) There exist $\lambda > -1$ and a λ measure μ on Ω such that

$$\mu(\varsigma_i) = c_i \text{ for } i = 1, \ldots, n. \tag{5}$$

b) If $\sum_{i=1}^n c_i = 1$, there exists a probability μ on $(\Omega, \mathcal{P}(\Omega))$ fulfilling (5).

c) The condition $\max_{i=1,\ldots,n} c_i = 1$ is equivalent to each λ measure μ on Ω fulfilling (5) being a probability with a singleton support.

Proof

a) Let $\max_{i=1,\ldots,n} c_i < 1$ and $\sum_{i=1}^n c_i \neq 1$, the remaining cases will be dealt with separately in b) and c). Denote $D = (-1, 0) \cup (0, +\infty)$ and define the function $g_\Omega : D \to \Re$ as follows:

$$(\forall x \in D)\ g_\Omega(x) = \frac{1}{x}\left(\prod_{i=1}^n (1 + x c_i) - 1\right) - 1 =$$
$$= \sum_{i=1}^n c_i + x \sum_{i_1=1}^{n-1} \sum_{i_2=i_1+1}^n c_{i_1} c_{i_2} + \cdots + x^{n-1} c_1 c_2 \cdots c_n - 1. \tag{6}$$

Then g_Ω is continuous on $(-1, 0) \cup (0, +\infty)$ and $\lim_{x\to 0} g_\Omega(x) = \sum_{i=1}^n c_i - 1$. Consequently, the condition $\sum_{i=1}^n c_i < 1$ implies $\lim_{x\to 0} g_\Omega(x) < 0$, which together with $\lim_{x\to+\infty} g_\Omega(x) = +\infty$ guarantees the existence of a zero point of g_Ω in $(0, +\infty)$, whereas $\sum_{i=1}^n c_i > 1$ implies $\lim_{x\to 0} g_\Omega(x) > 0$, which together with $\lim_{x\to -1} g_\Omega(x) = -\prod_{i=1}^n (1 - c_i) < 0$ guarantees the existence of a zero point of g_Ω in $(-1, 0)$. Let λ be such a point, and define:

$$\mu(\emptyset) = 0\ \&\ (\forall A = \{\varsigma_{i_1}, \ldots, \varsigma_{i_k}\} \in \mathcal{P}(\Omega))$$
$$\mu(A) = \frac{1}{\lambda}\left(\prod_{j=1}^k (1 + \lambda c_{i_j}) - 1\right). \tag{7}$$

For $A = \{\varsigma_i\}$, $i = 1, \ldots, n$, this directly yields (5). Combining (7) with (6) and with the fact that λ is a zero point of g_Ω leads to $\mu(\Omega) = 1$. To finish the proof of (1), consider disjoint sets $A = \{\varsigma_{i_1}, \ldots, \varsigma_{i_k}\}$ and $B = \{\varsigma_{i_{k+1}}, \ldots, \varsigma_{i_\ell}\}$. Then according to (7),

$$\mu(A \cup B) = \frac{1}{\lambda}(\prod_{j=1}^{\ell} (1 + \lambda c_{i_j}) - 1) =$$
$$= \frac{1}{\lambda}[(\prod_{j=1}^{k} (1 + \lambda c_{i_j}) - 1)(\prod_{j=k+1}^{\ell} (1 + \lambda c_{i_j}) - 1) +$$
$$+ \prod_{j=1}^{k} (1 + \lambda c_{i_j}) + \prod_{j=k+1}^{\ell} (1 + \lambda c_{i_j}) - 2] =$$
$$= \frac{1}{\lambda}(\prod_{j=1}^{k} (1 + \lambda c_{i_j}) - 1) + \frac{1}{\lambda}(\prod_{j=k+1}^{\ell} (1 + \lambda c_{i_j}) - 1) +$$
$$+ \lambda \frac{1}{\lambda}(\prod_{j=1}^{k} (1 + \lambda c_{i_j}) - 1) \frac{1}{\lambda}(\prod_{j=k+1}^{\ell} (1 + \lambda c_{i_j}) - 1) =$$
$$= \mu(A) + \mu(B) + \lambda\mu(A)\mu(B). \tag{8}$$

b) Since $\sum_{i=1}^n c_i = 1$, $(c_1, \ldots, c_n)$ is the vector of densities for some probability μ on $(\Omega, \mathcal{P}(\Omega))$. Then the fact that μ is a probability entails the validity of $\mu(\Omega) = 1$ and of $(\forall A, B \subset \Omega)\ A \cap B = \emptyset \to \mu(A \cup B) = \mu(A) + \mu(B)$, whereas the fact that $c_1, \ldots, c_n$ are densities implies the validity of (5).

c) If μ is a probability fulfilling (5) with a singleton support $\{\varsigma_{i_O}\}$, then $c_{i_O} = 1 = \max_{i=1,\ldots,n} c_i = 1$. On the other hand, let the condition $\max_{i=1,\ldots,n} c_i = 1$ and μ be a λ measure μ on Ω fulfilling (5). Denote $c_{i_O} = \max_{i=1,\ldots,n} c_i$. The existence of $i \neq i_0$ such that $c_i > 0$ would entail the contradiction

$$1 = \mu(\Omega) \geq \mu(\{\varsigma_{i_O}, \varsigma_i\}) =$$
$$= \mu(\{\varsigma_{i_O}\}) + \mu(\{\varsigma_i\}) + \lambda\mu(\{\varsigma_{i_O}\})\mu(\{\varsigma_i\}) =$$
$$= 1 + \mu(\{\varsigma_i\})(1 + \lambda) > 1. \tag{9}$$

Consequently, no such i exists, but $c_i = 0$ for each $i \neq i_0$. Then due to (1) and (5),

$$\mu(A) = \begin{cases} 1 & \text{if } \varsigma_{i_O} \in A, \\ 0 & \text{if } \varsigma_{i_O} \notin A, \end{cases} \tag{10}$$

which implies $\{\varsigma_{i_O}\}$ being the support of μ. To prove that μ is actually a probability, consider disjoint sets $A, B \subset \Omega$. Because at least one of A, B does not contain ς_{i_O}, (10) yields $\lambda\mu(A)\mu(B) = 0$, which in connection with (1) leads to $\mu(A \cup B) = \mu(A) + \mu(B)$. ■

Using the above proposition and its proof, a λ measure can be completely defined from the values $c_1, \dots, c_n$. Nevertheless, it is still necessary to define those values. Two principally different approaches are possible:

(i) The value c_i should measure some property of a particular classifier ς_i, $i = 1, \dots, n$. From the point of view of classification, the most desirable property of a classifier is its *generalization capability*, i.e., the capability to correctly classify unseen data. There are several possibilities how to measure that property [8], the most commonly used being the *accuracy* $\frac{1}{r}\,\mathrm{card}\{k : \mathbf{x}_k$ is correctly classified$\}$, which is an estimate of the probability of correct classification based on a random sample $(\mathbf{x}_1, \dots, \mathbf{x}_r)$ of unseen data.

(ii) The value c_i should measure some property of a particular class of classifiers to which ς_i belongs, such as the class of all SVMs based on a polynomial kernel of a particular order d, or the class of all SVMs based on an RBF kernel with a particular value of σ. Similarly to (i), the main concern is in the mean generalization capability of classifiers belonging to the class. It is typically measured using the method called *k-fold cross-validation*: The available data are divided into k parts and each part is used as a random sample to compute the accuracy of an SVM constructed using the remaining $k-1$ parts. The average accuracy of all k classifiers then serves as a measure of the mean generalization capability of SVMs from the class.

To combine at least two SVMs from Ω, μ must not have a singleton support. Then according to Propostion 5, $c_i \neq 1$ for $i = 1, \dots, n$. However, that condition can not be guaranteed if accuracies or their averages are used as c_i because they can happen to take the value 1. A simple remedy is to first multiply all accuracies with the same constant $\gamma \in (0, 1)$. In practical applications of the presented approach, the value $\gamma = \frac{1}{2}$ has been used, similarly to [18].

3. T-conorm system. For the choice of a t-conorm system, the following observations are useful:

- According to [7], using an Archimedean system is equivalent to using the Choquet integral with a transformed measure and a transformed integrand.

- Due to (2), $\diamond$ has to be nonconstant with respect to nonzero values of the second argument if the fuzzy measure μ should be able to have any influence on the integral.

- If integrating with respect to a fuzzy measure is viewed as a generalization of integrating with respect to a probability measure, the fuzzy integral should be viewed as the mean value of the integrand with respect to the fuzzy measure. To this end, it is necessary that the integral space be the same as the integrand space, hence $\underline{\perp} = \triangle$.

These observation suggest to restrict attention to non-Archimedean systems $\mathcal{F} = (\triangle, \perp, \triangle, \diamond)$, in which $\diamond$ is nonconstant with respect to nonzero values of the second argument. A recent publication [17] has proven that then $\triangle = \perp = \underline{\perp} = \vee$. A prominent example of such t-conorm systems are systems used for quasi-Sugeno integrals, in which $\diamond$ is in addition a t-norm. In that context, it is important that for $(\vee, \vee, \vee, \diamond)$, where $\diamond$ is a t-norm, necessary and sufficient to be a t-conorm system for integration is fulfilling the requirements (ii) and (iii) from Definition 2 [17].

The most easy way to find a comprehensive contingent of t-norms suitable to be used as $\diamond$ in quasi-Sugeno integrals is to investigate simultaneously whole *families of t-norms*, the individual members of which differ from each other through the value of some particular parameter. Many such families have been listed in [11]. For this paper, the *Aczél-Alsina family*, $(\wedge^{\varepsilon}_{\mathrm{AA}})_{\varepsilon\in(0,\infty)}$, the *Dombi family* $(\wedge^{\varepsilon}_{\mathrm{D}})_{\varepsilon\in(0,\infty)}$ and the *Frank family* $(\wedge^{\varepsilon}_{\mathrm{F}})_{\varepsilon\in(0,\infty)}$ have been selected, all of which have a parameter $\varepsilon \in (0, \infty)$. The following definitions are needed only for $x, y \in (0, 1)$ because for the remaining pairs (x, y), any t-norm $\diamond$ fulfils $0 \diamond x = x \diamond 0 = 0$ and $1 \diamond x = x \diamond 1 = x$:

$$x \wedge^{\varepsilon}_{\mathrm{AA}} y = e^{-\sqrt[\varepsilon]{(-\log x)^{\varepsilon} + (-\log y)^{\varepsilon}}}, \tag{11}$$

$$x \wedge^{\varepsilon}_{\mathrm{D}} y = \frac{1}{1 + \sqrt[\varepsilon]{\left(\frac{1-x}{x}\right) + \left(\frac{1-y}{y}\right)}}, \tag{12}$$

$$x \wedge^{\varepsilon}_{\mathrm{F}} y = \begin{cases} \log_{\varepsilon}\left(1 + \frac{(\varepsilon^x - 1)(\varepsilon^y - 1)}{\varepsilon - 1}\right) & \text{if } \varepsilon \neq 1, \\ xy & \text{if } \varepsilon = 1. \end{cases} \tag{13}$$

It is easy to prove that those three families fulfil the requirements (ii) and (iii) concerning $\diamond$ in Definition 2. In addition, the property $\lim_{\varepsilon\to\infty} x \wedge^{\varepsilon}_{\mathrm{AA}} y = \lim_{\varepsilon\to\infty} x \wedge^{\varepsilon}_{\mathrm{D}} y = \lim_{\varepsilon\to 0} = x \wedge y$ has been proven in [11].

4. Integrand. Although SVMs have the dichotomous output 1 or -1, their construction based on training data $(x_1, y_1), \dots, (x_p, y_p)$ actually employs the graded information $\sum_{k\in S} \alpha_k y_k \kappa(\mathbf{x}_k, \mathbf{x}) + b$, which is the normal coordinate of the classified point with respect to the central hyperplane of the margin between the two classes [2, 15]. Since that value can theoretically equal any real number, a nondecreasing transformation $f_1 : \Re \rightarrow (0, 1)$ has to be applied to it for an integrand representing the membership of $\mathbf{x}$ in the class with a positive value of that coordinate, and a nonincreasing transformation $f_2 : \Re \rightarrow (0, 1)$ has to be applied to it for an integrand representing the membership of $\mathbf{x}$ in the class with a negative value of that coordinate. Our implementation of the presented approach uses the transformations $f_1(t) = \tanh(at)$ for $t \geq 0$, $f_2(t) = -\tanh(at)$ for $t \leq 0$,where $a > 0$ is a constant, and $f_1, f_2 = 0$ else. To fix the value of a, we have added the requirement $f_1(1) = f_2(-1) = 0.95$, which in the context of SVMs concerns exactly the support vectors [2, 15].

4 A Neurophysiologic Application

The proposed method of classifier combining has already been used in two real-world applications, the first of which will now be very briefly recalled. It is a neurophysiologic application concerning research into the detection of driver's somnolence, which has been performed in collaboration of neurophysiologists and transportation scientists at the Czech Technical University Prague [4]. The ultimate objective of that research is to provide a knowledge base and detection algorithms for a system able to detect driver's somnolence, a cause of severe traffic accidents.

The used data consist of 762 pairs of spectra of EEG segments with mean duration approx 3.5 s recorded for 24 probands with acute sleep deprivation, using standard 19-channel EEG measurement. Each pair contains spectra for signals from channels T3 and O1, which have been considered most informative by the participating neurophysiologists. The spectra have been obtained from the EEG signals by means of the Burg filter of order 20, and contain spectral densities for frequencies 0-30 Hz by step 1 Hz. Each of the included EEG segments corresponds to one of the following states of the proband: relaxed *vigilance*, *mental activity*, in particular solving a part of the Raven test, and *somnolence*. Though this most naturally leads to a 3-class classification, most important from the point of view of the ultimate objective of the research is the entailed 2-class classification between somnolence and the other states. Hence, $y_k = 1$ for a segment corresponding to somnolence, whereas $y_k = -1$ for a segment corresponding to vigilance or mental activity. The correct assignment of states to the 762 available segments was provided by an expert neurophysiologist, based on a visual inspection of the segments and their spectra.

The quality of the combined classifiers was tested, using 10-fold cross validation, for 63 different t-norms in the quasi-Sugeno integral, taken from the Aczél-Alsina, Dombi and Frank families. The results confirm the extreme difficulty of this real-world classification task: the mean error rate was approximately 40% and varied substantially between the individual test sets of the cross validation – for some of them it was only slightly over 10%, whereas for others it was worse than that of a random classifier. From the point of view of the topic of this paper, i.e., using the quasi-Sugeno integral to combine classifiers, very important is that the results are nearly insensitive to the choice of the t-norm in the integral, more precisely to the choice of the value of the parameter ε in the three families. The lowest error rates were achieved with $\varepsilon = 1.3 - 1.6$ for the Aczél-Alsina family, $\varepsilon = 0.5 - 0.8$ for the Dombi family, and $\varepsilon = 0.03 - 0.25$ for the Frank family, but in all those cases, the relative improvement compared to the traditional Sugeno integral is only about 1%. For illustration, Figure 1 shows the error rate for the Dombi family.

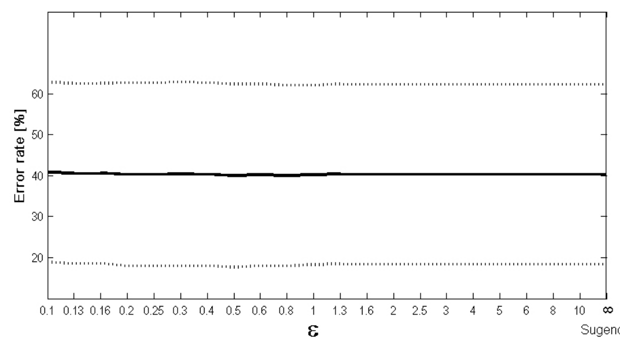

Figure 1: Dependence of the error rate on ε for the Dombi family (solid line: mean, dotted lines: mean $\pm$ standard deviation)

5 Conclusions

This paper continues a research into an approach to combining classifiers based on the t-conorm integral, in particular on the quasi-Sugeno integral, the first results of which have been recently published in [17]. That approach has been further theoretically developed, and elaborated specifically for support vector machines. The result of its theoretical development is Propostion 5, which analyses situations that can occur when a λ measure is constructed. Such a proposition has not been available in [17]. As far as the elaboration for support vector machines is concerned, a crucial role plays the choice of integrand as a transformation of the normal coordinate with respect to the central hyperplane of a margin. The experience from a first real-world application has most importantly brought the observation that the result of the approach is nearly insensitive to the choice of the t-norm in the quasi-Sugeno integral. It is worth mentioning that a similar observation has been done also in the case of a more recent application of the approach to data described in [13] and concerning catalytic materials, and in the case of the well known benchmark of Fisher iris data [5].

Acknowledgments

The research reported in this paper has been supported by the grant No. ME701, ”Building Neuroinformation Bases, and Extracting Knowledge from them”, of the Czech Ministry for Education, and partially supported by the Institutional Research Plan AVOZ10300504.

References

[1] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, 1984.

[2] C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Knowledge Disocvery and Data Mining*, 2:121–167, 1998.

[3] G. Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5:131–296, 1954.

[4] J. Faber, M. Novák, P. Svoboda, and V. Tatarinov. Electrical brain wave analysis during hypnagogium. *Neural Network World*, 13:41–54, 2003.

[5] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[6] M. Grabisch, T. Murofushi, and M. Sugeno. Fuzzy measure of fuzzy events defined by fuzzy integrals. *Fuzzy Sets and Systems*, 50:293–313, 1992.

[7] M. Grabisch, H.T. Nguyen, and E.A. Walker. *Fundamentals of Uncertainty Calculi with Applications to Fuzzy Inference*. Kluwer Academic Publishers, Dordrecht, 1995.

[8] D.J. Hand. *Construction and Assessment of Classification Rules*. John Wiley and Sons, New York, 1997.

[9] T. Hasti, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Verlag, New York, 2001.

[10] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:226–239, 1998.

[11] E.P. Klement, R. Mesiar, and E. Pap. *Triangular Norms*. Kluwer Academic Publishers, Dordrecht, 2000.

[12] L.I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley and Sons, New York, 2004.

[13] S. Moehmel, N. Steinfeldt, S. Endgelschalt, M. Holeňa, S. Kolf, U. Dingerdissen, D. Wolf, R. Weber, and M. Bewersdorf. New catalytic materials for the high-temperature synthesis of hydrocyanic acid from methane and ammonia by high-throughput approach. Accepted for publication in Applied Catalysis A: General.

[14] J. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Francisco, 1992.

[15] B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.

[16] M. Sugeno. *Theory of Fuzzy Integrals and Its Application*. PhD thesis, Tokyo Institute of Technology, 1974.

[17] D. Štefka and M. Holeňa. The use of fuzzy t-conorm integral for combining classifiers. In *Proceedings of the Ninth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. Springer Verlag, Berlin, 2007.

[18] D. Wang, J.M. Keller, C.A. Carson, K.K. Mc Adoo-Edwards, and C.W. Bailey. Use of fuzzy-logic-inspired features fo improve bacterial recognition through classifier fusion. *IEEE Transactions on Systems, Man and Cybernetics. Part B Cybernetics*, 28:583–591, 1998.