Classifier Aggregation using Fuzzy Integral based on Interaction-Sensitive Fuzzy Measures

David Štefka

28. March 2013

Martin Holeňa

Classifier Combining

- classification – predict to which class a given pattern belongs
- classifier combining/aggregation/fusion/selection/...
  - create a team of classifiers and aggregate their predictions
  - better generalization properties
  - lower error rate
  - better robustness
  - less sensitive to overfitting
  - the resulting system behaves as a single classifier
  - no generally accepted unifying theory
  - how does it work? Bias/variance decomposition (variance is reduced), large margin classifiers (large margin $\rightarrow$ better generalization)

Classifier Team Design

- motivation: induce *diversity* to the team
- sampling from the training set (bagging, boosting)
- partitioning the feature space (divide&conquer, mixture of experts)
- using different combinations of features (multiple feature subset, attribute bagging)
- multi-model approaches (e.g., k-NN, neural net, decision tree, and SVM)
- changing parameters of a model (3-NN, 5-NN, 10-NN; neural net topology)
- output coding (error correcting output coding)
- hybrid methods (random forests)

Classification Confidence

- motivation: measure the degree of reliability of a prediction
- *static*
  - global accuracy, precision, sensitivity, . . .
- *dynamic*
  - local accuracy
  - local match
  - methods based on d.o.c.
  - statistical methods - transduction
  - model-specific methods

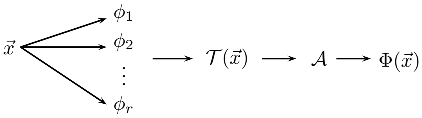
Aggregation

- classifier selection (static/dynamic classifier selection, mixture of experts)
- crisp classifiers - voting, behavior knowledge space
- class ranking methods - Borda count
- soft classifiers - artihmetic approaches (mean, median, min, max), probabilistic approaches (product rule, Dempster-Shafer theory), fuzzy logic (fuzzy integral, decision templates)
- second level classifiers - stacking

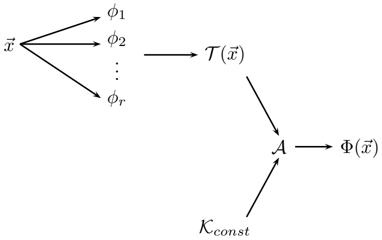
Dynamic Classifier Systems

- framework of classifier combining with classification confidence
- $\mathcal{S} = (\mathcal{T}, \mathcal{K}, \mathcal{A})$ – classifier system
- $\mathcal{T} = (\phi_1, \ldots, \phi_r)$ – classifiers
- $\mathcal{K} = (\kappa_{\phi_1}, \ldots, \kappa_{\phi_r})$ – confidence measures
- $\mathcal{A}$ – aggregator
- 3 types of classifier systems
  - confidence-free
  - static
  - dynamic

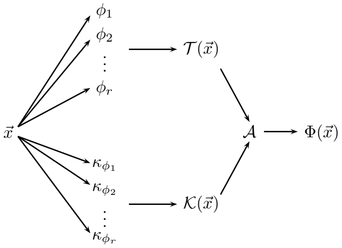
Types of classifier systems



(a) Confidence-free



(b) Static



(c) Dynamic

Classifier Aggregation

- prediction

$$\mathcal{T}(\vec{x}) = \begin{pmatrix} \phi_1(\vec{x}) \\ \phi_2(\vec{x}) \\ \vdots \\ \phi_r(\vec{x}) \end{pmatrix} = \begin{pmatrix} \gamma_{11}(\vec{x}) & \gamma_{12}(\vec{x}) & \dots & \gamma_{1N}(\vec{x}) \\ \gamma_{21}(\vec{x}) & \gamma_{22}(\vec{x}) & \dots & \gamma_{2N}(\vec{x}) \\ & & \ddots & \\ \gamma_{r1}(\vec{x}) & \gamma_{r2}(\vec{x}) & \dots & \gamma_{rN}(\vec{x}) \end{pmatrix}$$

$$\gamma_{ij}(\vec{x}) = \text{degree of classification to class } C_j \text{ given by } \phi_i$$

- confidence

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

$$\kappa_{\phi_i}(\vec{x}) = \text{confidence of } \phi_i \text{ on } \vec{x}$$

- usually, aggregate j-th column of $\mathcal{T}(\vec{x})$ by an aggregation operator, parametrized by $\mathcal{K}(\vec{x})$

Information Fusion

- $(X_1, \ldots, X_N)$ – information sources (sensors, experts, etc.)
- $(a_1, \ldots, a_N) \in D^N$– outputs in domain D, e.g. $D = \mathcal{R}$
- $\mathbb{C} : \ D^N \to D$ – aggregation operator
- $\mathbb{C}(a_1, \ldots, a_N)$ – aggregated value (consensus)
- arithmetic mean, weighted mean, median, minimum, maximum, ...

Desired Properties

- unanimity
  $\forall a : \mathbb{C}(a, \ldots, a) = a$
- monotonicity
  $\forall i : a_i \geq a'_i \Rightarrow \mathbb{C}(a_1, \ldots, a_N) \geq \mathbb{C}(a'_1, \ldots, a'_N)$
- (unanimity) + (monotonicity) $\Rightarrow$ internality
  $\min_i a_i \leq \mathbb{C}(a_1, \ldots, a_N) \leq \max_i a_i$
- symmetry (no source is distingushable)
  $\forall \pi \in \Pi_{1,\ldots,N} : \mathbb{C}(a_1, \ldots, a_N) = \mathbb{C}(a_{\pi(1)}, \ldots, a_{\pi(N)})$
- robustness (influence of outliers) - arithmetic mean vs. median
- applicability – numeric / ordinal / nominal domains

Weighted Mean

- $WM_p(a_1, \ldots, a_N) = \sum_i p_i a_i$
- weighting vector: $\mathbf{p} = (p_1, \ldots, p_N) \in [0,1]^N, \ \sum_i p_i = 1$
- p_i – importance (reliability) of i-th source
- properties
  - special case – arithmetic mean ($p_i = 1/N$)
  - not symmetric
  - dictatorship of the i-th source ($p_i = 1, \ p_j = 0 \ j \neq i$)
  - unbounded influence of outliers

Ordered Weighted Average (OWA)

- $OWA_w(a_1, \ldots, a_N) = \sum_i w_i a_{<i>}$
- weighting vector **w**, $(\cdot)$ indicating nondecreasing permutation, i.e. $a_{<i>} \geq a_{<i-1>}$
- w_i – importance of i-th largest output
- properties
  - can reduce (or ignore) extreme values, e.g. $\mathbf{w} = (0, 1/3, 1/3, 1/3, 0)$ – commitee
  - special cases – minimum, maximum, median, arithmetic mean
  - symmetric

Fuzzy Measure

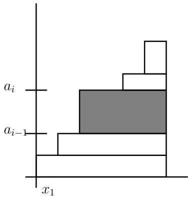
- $\mu : \mathcal{P}(\mathcal{U}) \rightarrow [0,1]$ is called a fuzzy measure on $\mathcal{U}$ iff:
  1. (boundary condition) $\mu(\emptyset) = 0$, $\mu(\mathcal{U}) = 1$
  2. (monotonicity) $A \subseteq B \Rightarrow \mu(A) \leq \mu(B)$
- generalization of additive measures (probability)
- can model interaction between the elements

  example: 3 subjects (math, physics, literature); $\mu(\emptyset) = 0$, $\mu(M) = 0.45$, $\mu(P) = 0.45$, $\mu(L) = 0.3$, $\mu(M,L) = 0.9$, $\mu(P,L) = 0.9$, $\mu(M,P) = 0.5$, $\mu(M,P,L) = 1$
- classifier aggregation: aggregate the integrand (predictions of the classifiers) with respect to the fuzzy measure (represents the confidence)
- no general definition of fuzzy integral; Choquet and Sugeno used most often

Choquet Integral

i	support; $A_{<i>}$	d.o.c.-level; $f_{<i>}$	measure of support; $\mu(A_{<i>})$
4	ϕ_3	0.9	0.1
3	ϕ_3, ϕ_4	0.4	0.3
2	ϕ_1, ϕ_3, ϕ_4	0.3	0.7
1	$\phi_1, \phi_2, \phi_3, \phi_4$	0.2	1
0		0	

$$\int_C f\, d\mu = \sum_{i=1}^{r} (f_{<i>} - f_{<i-1>})\mu(A_{<i>})$$
$$= 0.5 \cdot 0.1 + 0.1 \cdot 0.3 + 0.1 \cdot 0.7 + 0.2 \cdot 1 = 0.35$$



Choquet Integral ctnd

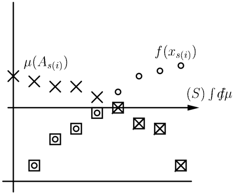
- for additive measures, Choquet integral conincides with Lebesgue integral
- satisfies unanimity, monotonicity, internality (i.e., it is a proper aggregation operator)
- generalizes weighted mean, OWA, WOWA

Sugeno Integral

i	support; $A_{<i>}$	d.o.c.-level; $f_{<i>}$	measure of support; $\mu(A_{<i>})$
4	ϕ_3	0.9	0.1
3	ϕ_3, ϕ_4	0.4	0.3
2	ϕ_1, ϕ_3, ϕ_4	0.3	0.7
1	$\phi_1, \phi_2, \phi_3, \phi_4$	0.2	1
0		0	

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

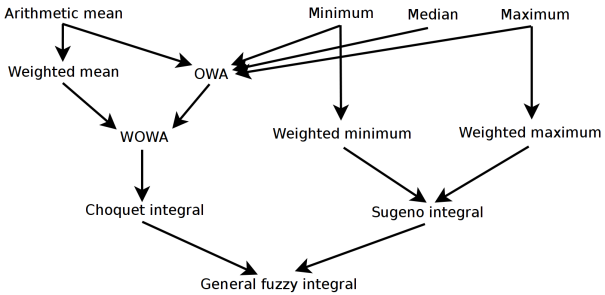
$$= max(0.1, 0.3, 0.3, 0.2) = 0.3$$



Sugeno Integral ctnd

- satisfies unanimity, monotonicity, internality (i.e., it is a proper aggregation operator)
- generalizes weighted minimum and maximum

Aggregation Operators - summary



Fuzzy Integral

- aggregate the integrand w.r.t. fuzzy measure
- integrand $\sim$ degrees of classification (d.o.c.) to C_j given by $\phi_1, \ldots, \phi_r$
- fuzzy measure $\sim$ confidences of the individual classifiers
- integral $\sim$ aggregated d.o.c. to class C_j

Fuzzy Measure

- $\mu : \mathcal{P}(X) \to [0,1]$ is called a fuzzy measure on X iff:
  1. (boundary condition) $\mu(\emptyset) = 0$, $\mu(X) = 1$
  2. (monotonicity) $A \subseteq B \Rightarrow \mu(A) \leq \mu(B)$
- generalization of additive measures (probability)
- can model interaction between the elements

  example: 3 subjects (math, physics, literature); $\mu(\emptyset) = 0$, $\mu(M) = 0.45$, $\mu(P) = 0.45$, $\mu(L) = 0.3$, $\mu(M, L) = 0.9$, $\mu(P, L) = 0.9$, $\mu(M, P) = 0.5$, $\mu(M, P, L) = 1$
- hard to define (needs $2^N - 2$ parameters)
- additive measures need only $N - 1$ parameters (for the singletons) - fuzzy densities $\mu(\phi_i)$

Common Fuzzy Measures

- additive: $\mu(A \cup B) = \mu(A) + \mu(B)$ for disjoint A, B
  - correspond to probabilistic measures
- symmetric: $|A| = |B| \Rightarrow \mu(A) = \mu(B)$
  - $\mu(A)$ depends only on the number of elements in A
  - leads to confidence-free aggregation
- $\perp$-decomposable: $\mu(A \cup B) = \mu(A) \perp \mu(B)$ for disjoint A, B
  - special case: Sugeno λ-measure (used most often in classifier aggregation using FI); $\mu(A \cup B) = \mu(A) + \mu(B) + \lambda\mu(A)\mu(B)$
  - $\mu(A \cup B)$ fully determined by $\mu(A), \mu(B), \perp$
- neither of these can model interactions between the classifiers

Interaction-Sensitive Fuzzy Measures

- motivation: model the confidence of a set of classifiers, but take mutual classifier similarities ($\sim$ interactions) into account
- similar classifiers: small increase in the measure
- different classifiers: big increase in the measure
- diversity of the classifier team is taken into account in the aggregation process (not processed a priori)
- not limited to classifier aggregation only

Induced Interaction-Sensitive Fuzzy Measure (I-ISFM)

- at each step, classifier $\phi_{<i>}$ is added to a set of classifiers $(\phi_{<i+1>}), \ldots, \phi_{<r>})$
- increase of the measure is controlled by the similarity

$$
\begin{aligned}
\mu(\emptyset) &= 0 \\
\mu(A_{<r>}) &= \mu(\{\phi_{<r>}\}) = \kappa_{<r>} \\
\mu(A_{<i>}) &= \mu(\{\phi_{<i>}, \ldots, \phi_{<r>}\}) = \\
&= \mu(A_{<i+1>}) + [1 - \max_{k=i+1}^{r} S(\phi_{<i>}, \phi_{<k>})]\kappa_{<i>} \\
&\text{for } i = r-1, \ldots, 1,
\end{aligned}
$$

- I-ISFM: μ normalized to $[0, 1]$
- theoretical weakness: tightly connected to the ordering $< \cdot >$ induced by f

Global Interaction-Sensitive Fuzzy Measure (G-ISFM)

- fuzzy measure on the whole universe; regardless of the integrand
- take the classifier confidences and transform them into new fuzzy densities

$$\mu(\phi_k) = \kappa_k \rightsquigarrow \widetilde{\mu}(\phi_k)$$

- classifiers are sorted w.r.t. confidences $[\cdot]$
- with decreasing confidence, the similarity to elements with higher confidence is taken into account

$$\widetilde{\mu}(\phi_{[k]}) = \kappa_{[k]}(1 - \max_{j=k+1}^{r} s_{[k],[j]}), \; k = 1, \dots, r$$

- use $\widetilde{\mu}(\phi_{[k]})$ to build an additive measure

Modified Hüllermeier Measure (MHM)

- Cho-k-NN: use similarities of neighbors in k-NN classifier
- base measure ν (e.g., additive, based on the confidences)
- use diversity of a set of classifiers to adjust the base measure

$$div(A) = \frac{2}{|A|^2 - |A|} \sum_{u_i, u_j \in A; j<i} (1 - s_{i,j}) \in [0,1]$$

$$rdiv(A) = \frac{2div(A)}{\max(1 - s_{i,j})} - 1 \in [-1,1]$$

$$\mu_h(A) = \nu(A)(1 + \alpha rdiv(A)),\ \alpha \geq 0$$

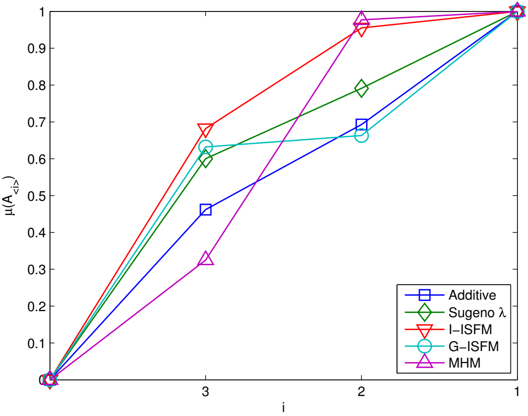
- not necessarilly monotone
  - enforce monotonicity using $\mu_h(A) = \max_{B \subseteq A} \mu_h(B)$ is practically impossible
  - use the idea from I-ISFM: compute μ_h only for the r values actually needed for the integration, i.e., sets $A_{<i>}$

Example - similar classifiers

$$\mathcal{T}_{*,j}(\vec{x}) = [0.5, 0.4, 0.8]^T$$
$$\mathcal{K}(\vec{x}) = [0.3, 0.4, 0.6]^T$$
$$(s_{i,j}) = \begin{pmatrix} 1 & 0.9 & 0.2 \\ 0.9 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{pmatrix}$$

i	support $A_{<i>}$	d.o.c.-level $f_{<i>}$	$\mu(A_{<i>})$ additive	Sugeno λ	I-ISFM	G-ISFM	MHM
3	ϕ_3	0.8	0.462	0.6	0.682	0.632	0.325
2	ϕ_1, ϕ_3	0.5	0.693	0.791	0.955	0.663	0.977
1	ϕ_1, ϕ_2, ϕ_3	0.4	1	1	1	1	1

Example - similar classifiers



Similar classifiers

$$\mathcal{T}_{*,j}(\vec{x}) = [0.5, 0.4, 0.8]^T$$

$$\mathcal{K}(\vec{x}) = [0.3, 0.4, 0.6]^T$$

$$(s_{i,j}) = \begin{pmatrix} 1 & 0.9 & 0.2 \\ 0.9 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{pmatrix}$$

Example - dissimilar classifiers

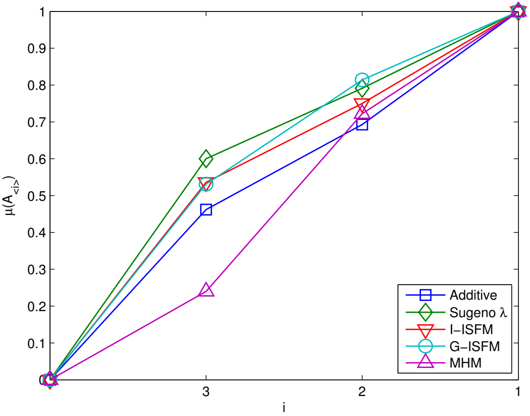
$$\mathcal{T}_{*,j}(\vec{x}) = [0.5, 0.4, 0.8]^T$$

$$\mathcal{K}(\vec{x}) = [0.3, 0.4, 0.6]^T$$

$$(s_{i,j}) = \begin{pmatrix} 1 & 0.3 & 0.2 \\ 0.3 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{pmatrix}$$

i	support $A_{<i>}$	d.o.c.-level $f_{<i>}$	$\mu(A_{<i>})$ additive	Sugeno λ	I-ISFM	G-ISFM	MHM
3	ϕ_3	0.8	0.462	0.6	0.536	0.531	0.240
2	ϕ_1, ϕ_3	0.5	0.693	0.791	0.75	0.814	0.722
1	ϕ_1, ϕ_2, ϕ_3	0.4	1	1	1	1	1

Example - dissimilar classifiers



Dissimilar classifiers

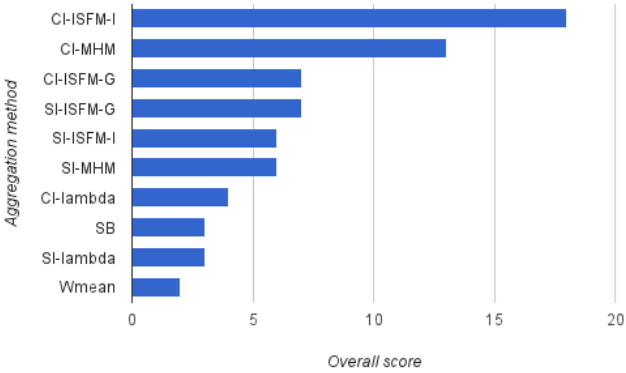
$$\mathcal{T}_{*,j}(\vec{x}) = [0.5, 0.4, 0.8]^T$$
$$\mathcal{K}(\vec{x}) = [0.3, 0.4, 0.6]^T$$

$$(s_{i,j}) = \begin{pmatrix} 1 & 0.3 & 0.2 \\ 0.3 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{pmatrix}$$

Experiments

- compare non-interaction sensitive measures (additive, Sugeno λ-measure) to ISFM (I-ISFM, G-ISFM, MHM)
- 3 different classifier systems (Random Forest, k-NN ensemble, QDC ensemble)
- 23 datasets
- Choquet/Sugeno integral with Sugeno λ-measure and ISFM
- reference: single best, weighted mean ($\sim$ additive measure)

Experimental results



Number of datasets (out of 69), for which the aggregator obtained the best results among all aggregators.

Experimental results

↓ superior to → (out of 69)	SB	WMean	CI				SI				all
			λ	I-ISFM	G-ISFM	MHM	λ	I-ISFM	G-ISFM	MHM	
SB	-	32 (4)	19 (1)	9	9	11	18 (3)	9	12	12 (1)	3
WMean	37 (16)	-	15 (2)	5	15	7	18 (2)	7	22 (2)	10 (1)	2
CI-λ	50 (18)	54 (6)	-	9	20	13	41	13	28 (1)	15 (1)	4
CI-I-ISFM	60 (23)	64 (19)	60 (7)	-	45 (2)	38	57 (7)	45 (1)	49 (7)	47 (1)	18
CI-G-ISFM	61 (24)	54 (18)	49 (7)	24	-	28 (1)	53 (10)	32 (1)	48	36 (2)	7
CI-MHM	58 (24)	62 (17)	56 (7)	31	43 (2)	-	57 (7)	42 (1)	48 (6)	41 (1)	13
SI-λ	51 (17)	51 (6)	30	12	16	12	-	13	28 (1)	18 (1)	3
SI-I-ISFM	61 (24)	62 (17)	56 (6)	27	39 (2)	28	56 (8)	-	48 (4)	39 (1)	6
SI-G-ISFM	58 (23)	47 (14)	41 (5)	20	21	21	41 (7)	23 (1)	-	27	7
SI-MHM	58 (23)	59 (13)	54 (6)	22	33 (2)	28	51 (8)	30 (1)	43 (4)	-	6

Number of datasets (out of 69), on which aggregator i obtained better results than aggregator j, including significant improvements in parentheses.

Experimental results

- ISFMs generally outperform traditional fuzzy measures (often significantly)
- CI obtained better results than SI
- I-ISFM and MHM slightly superior to G-ISFM

Conclusions

- dynamic classifier systems aggregated using fuzzy integral
- traditional fuzzy measures (additive, symmetric, $\perp$-decomposable) do not take classifier similarities into account
- ISFM: use classifier similarities in the fuzzy measure to further improve the fuzzy integral-based aggregation
- three novel fuzzy measures: I-ISFM, G-ISFM, MHM
- diversity is processed directly in the aggregation
- fast evaluation
- not limited to classifier aggregation only
- experimental results: ISFMs outperform traditional fuzzy measures

Thank you for your attention

David Štefka
stefka@insophy.cz

Classifier Aggregation
using Fuzzy Integral based on
Interaction-Sensitive Fuzzy Measures