Enhancing Testing-Time Robustness for Trusted Multi-View Classification in the Wild

Supplementary Material

A. Proof of Propositions

Proposition 1. *$\Lambda(\kappa^v)$ is an (α, δ)-risk-controlling discovery (RCD) function.*

Proof:

1. Problem Setup and Objective

We aim to prove that the discovery function $\Lambda(\kappa^v)$ satisfies the properties of an (α, δ)-risk-controlling discovery (RCD) function. Specifically, we want to show:

$$\mathbb{P}\left(\mathrm{FDR}\left(\Lambda\left(\kappa^v\right)\right) \leq \alpha\right) \geq 1-\delta,$$

where the false discovery rate (FDR) is defined as:

$$\mathrm{FDR}\left(\Lambda\left(\kappa^v\right)\right) := \mathbb{E}\left[\frac{\sum_{j \in D_{test}} \mathbb{1}\left\{j \in S^v, u_j^v > \kappa^v\right\}}{\max\left\{1, |S^v|\right\}}\right].$$

2. Definitions and Threshold Construction

Given the validation set D_{val} containing L samples, their uncertainty scores $u_1^v, \ldots, u_L^v$ are exchangeable. Sorting these scores gives:

$$u_{(1)}^v \leq u_{(2)}^v \leq \ldots \leq u_{(L)}^v.$$

According to the Equation (8), the threshold κ^v is defined as the $(1-\alpha)$-quantile of the ordered validation scores:

$$\kappa^v = u_{(\lceil (1-\alpha)(1+L) \rceil)}^v.$$

This definition ensures that at most $\alpha \cdot (L+1)$ samples from the validation set exceed κ^v.

For any test sample $j \in D_{test}$, the p-value is defined as:

$$p\text{-value}\left(\boldsymbol{e}_j^v\right) = \frac{1 + \sum_{l \in D_{val}} \mathbb{1}\left\{u_j^v \geq u_l^v\right\}}{L+1}.$$

This p-value represents the fraction of validation uncertainty scores less than or equal to u_j^v. If $u_j^v > \kappa^v$, the null hypothesis $u_j^v \leq \kappa^v$ is rejected, classifying j as unreliable, and then we have:

$$\begin{aligned} p\text{-value}\left(\boldsymbol{e}_j^v\right) &= \frac{1 + \sum_{l \in D_{val}} \mathbb{1}\left\{u_j^v \geq u_l^v, u_j^v > \kappa^v\right\}}{L+1} \\ &= \frac{1 + \sum_{l \in D_{val}} \mathbb{1}\left\{u_l^v > \kappa^v\right\}}{L+1} \\ &= \frac{1 + L - \lceil (1-\alpha)(L+1) \rceil}{L+1} \\ &\leq \alpha. \end{aligned}$$

3. Marginal Validity

The p-values are marginally valid because κ^v is constructed from the $(1-\alpha)$-quantile of D_{cal}. Specifically:

$$\mathbb{P}\left(u_j^v > \kappa^v\right) \leq \alpha \Rightarrow \mathbb{P}\left(p\text{-value}\left(\boldsymbol{e}_j^v\right) \leq \alpha\right) \leq \alpha.$$

Thus, the probability of incorrectly classifying a test sample as unreliable (a false discovery) is bounded by α.

4. False Discovery Rate (FDR)

False positives (FP) are defined as:

$$\mathrm{FP} = \sum\nolimits_{j \in D_{test}} \mathbb{1}\left\{j \in S^v, u_j^v > \kappa^v\right\}.$$

Since $\mathbb{P}\left(u_j^v > \kappa^v\right) \leq \alpha$, the expected number of false positives (FP) is bounded by:

$$\mathbb{E}\left[\mathrm{FP}\right] \leq \alpha \cdot nulls,$$

where $nulls$ is the number of null hypotheses in D_{test}. The total discoveries $|S^v|$ include both false positives (FP) and true positives (TP). Thus $|S^v| = \mathrm{FP} + \mathrm{TP}$.

The FDR is defined as:

$$\mathrm{FDR}\left(\Lambda\left(\kappa^v\right)\right) = \mathbb{E}\left[\frac{\mathrm{FP}}{\max\left\{1, |S^v|\right\}}\right].$$

since $S^v \geq \mathrm{FP}$, and the numerator FP is bounded by marginal validity, we have:

$$\mathrm{FDR} \leq \alpha.$$

5. High-Probability Control of FDR

To ensure that FDR is controlled with high probability $1-\delta$, consider the variability in κ^v:

- *The threshold κ^v is determined from the validation set D_{val}, whose scores are exchangeable.*
- *Using concentration inequalities (e.g., Hoeffding's inequality or Dvoretzky-Kiefer-Wolfowitz (DKW) inequality), the deviation of κ^v from its expectation is bounded with probability $1-\delta$.*

Specifically, using the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality, we can bound the deviation of the empirical CDF from the true CDF. Let $F(u)$ be the true cumulative distribution function (CDF) of the uncertainty scores. For a sample of size L, the DKW inequality states:

$$\mathbb{P}\left(\sup_u \left|\widehat{F}_L\left(u\right) - F\left(u\right)\right| > \epsilon\right) \leq 2e^{-2L\epsilon^2},$$

where $\widehat{F}_L\left(u\right)$ is the empirical CDF.

Applying this to κ^v, the probability that the estimated $(1-\alpha)$-quantile deviates from the true $(1-\alpha)$-quantile by more than ϵ is bounded by:

$$\mathbb{P}\left(\left|\kappa^v - F^{-1}(1-\alpha)\right| > \epsilon\right) \leq 2e^{-2L\epsilon^2} \leq \delta,$$

where $\delta = 2e^{-2L\epsilon^2}$. This guarantees that κ^v is concentrated around the true $(1-\alpha)$-quantile with high probability, provided the validation set is sufficiently large.

Thus, the probability of FDR *exceeding α is:*

$$\mathbb{P}\left(\mathrm{FDR}\left(\Lambda\left(\kappa^v\right)\right) > \alpha\right) \leq \delta.$$

Equivalently:

$$\mathbb{P}\left(\mathrm{FDR}\left(\Lambda\left(\kappa^v\right)\right) \leq \alpha\right) \geq 1-\delta.$$

6. Special Case: $\alpha < 1/(L+1)$

when $\alpha < 1/(L+1)$ (a rare situation in the wild), the threshold κ^v is undefined (or can be defined as infinity) because the validation set is too small to reliably estimate the $(1-\alpha)$-quantile. This highlights the need for: 1) A sufficiently large validation set such that $(L+1) > 1/\alpha$; 2) A practical lower bound on α based on the size of D_{val}.

Note, when κ^v is defined as infinity, indicating that no test samples will satisfy $u_j^v > \kappa^v$. This effectively means no discoveries will be made, which trivially satisfies FDR control but is not useful in practice.

To sum up, $\Lambda\left(\kappa^v\right)$ is an (α,δ)-risk-controlling discovery (RCD) function.

□

Proposition 2. *Applying evidence filtering improves the classification performance of TMVC fusion strategies (e.g., BCF, A-CBF, or ABF), under the assumption that unreliable evidence provides incorrect information in the wild.*

Proof: *For $k = 1, \ldots, K$, let t denote the index of the ground-truth class, and $\{e_{ik}^o\}_{k=1}^K$ be the initial evidence from some views. Consider an abnormal view (i.e., noisy or corrupted) producing unreliable evidence $\{e_{ik}^v\}_{k=1}^K$, where $\tilde{k} = \arg\max_k \{e_{ik}^v\}_{k=1}^K, \tilde{k} \neq t$. Assume this unreliable evidence is strong enough to change the predicted label of the initial evidence to $\tilde{k}$, leading to incorrect classification. Let $\{e_{ik}\}_{k=1}^K$ and $\{\bar{e}_{ik}\}_{k=1}^K$ be the evidence after fusion of $\{e_{ik}^v\}_{k=1}^K$ with the initial evidence $\{e_{ik}^o\}_{k=1}^K$, with and without filtering, respectively. Taking A-CBF as an example, the prediction probabilities for class t under the two scenarios are given as follows:*

With filtering:

$$p_{it} = \frac{e_{it}}{\sum_{k=1}^K e_{ik} + 1} = \frac{e_{it}^o + \alpha \cdot e_{it}^v}{\sum_{k=1}^K \left(e_{ik}^o + \alpha \cdot e_{ik}^v\right) + 1},$$

Without filtering:

$$\bar{p}_{it} = \frac{\bar{e}_{it}}{\sum_{k=1}^K \bar{e}_{ik} + 1} = \frac{e_{it}^o + e_{it}^v}{\sum_{k=1}^K \left(e_{ik}^o + e_{ik}^v\right) + 1}.$$

Similarly, the prediction probabilities for the incorrect class $\tilde{k}$ are:

$$p_{i\tilde{k}} = \frac{e_{i\tilde{k}}}{\sum_{k=1}^K e_{ik} + 1} = \frac{e_{i\tilde{k}}^o + \alpha \cdot e_{i\tilde{k}}^v}{\sum_{k=1}^K \left(e_{ik}^o + \alpha \cdot e_{ik}^v\right) + 1},$$

$$\bar{p}_{i\tilde{k}} = \frac{\bar{e}_{i\tilde{k}}}{\sum_{k=1}^K \bar{e}_{ik} + 1} = \frac{e_{i\tilde{k}}^o + e_{i\tilde{k}}^v}{\sum_{k=1}^K \left(e_{ik}^o + e_{ik}^v\right) + 1}.$$

The difference between the probabilities for t and $\tilde{k}$ is:

$$\Delta p = p_{it} - p_{i\tilde{k}} = \frac{\left(e_{it}^o + \alpha \cdot e_{it}^v\right) - \left(e_{i\tilde{k}}^o + \alpha \cdot e_{i\tilde{k}}^v\right)}{\sum_{k=1}^K \left(e_{ik}^o + \alpha \cdot e_{ik}^v\right) + 1},$$

$$\Delta \bar{p} = \bar{p}_{it} - \bar{p}_{i\tilde{k}} = \frac{\left(e_{it}^o + e_{it}^v\right) - \left(e_{i\tilde{k}}^o + e_{i\tilde{k}}^v\right)}{\sum_{k=1}^K \left(e_{ik}^o + e_{ik}^v\right) + 1}.$$

Filtering scales the unreliable evidence by α, so the numerators satisfy:

$$\left(e_{it}^o + \alpha \cdot e_{it}^v\right) - \left(e_{i\tilde{k}}^o + \alpha \cdot e_{i\tilde{k}}^v\right) > \left(e_{it}^o + e_{it}^v\right) - \left(e_{i\tilde{k}}^o + e_{i\tilde{k}}^v\right).$$

Filtering also reduces the total contribution of unreliable evidence in the denominator:

$$\sum_{k=1}^K \left(e_{ik}^o + \alpha \cdot e_{ik}^v\right) + 1 < \sum_{k=1}^K \left(e_{ik}^o + e_{ik}^v\right) + 1.$$

Both the numerator and denominator changes lead to the inequality: $\Delta p > \Delta \bar{p}$. Similarly, for both ABF and BCF fusion strategies, we also observe that $\Delta p > \Delta \bar{p}$ because filtering reduces the impact of unreliable evidence, denoted as $\alpha \cdot e_{ik}^v$. This results in the ground-truth class t being more likely to have the highest prediction probability, i.e., $p_{it} > p_{i\tilde{k}}, \forall \tilde{k} \neq t$. As a consequence, the classification error rate is reduced, leading to an overall improvement in performance. □

Proposition 3. *The overall uncertainty of multi-view results generated by TMVC fusion strategies with evidence filtering for unreliable evidence will exhibit greater uncertainty than fusion without evidence filtering.*

Proof: *For $k = 1, \ldots, K$, let $\{e_{ik}^o\}_{k=1}^K$ be the initial evidence from some views. Consider an abnormal view with unreliable evidence $\{e_{ik}^v\}_{k=1}^K$, let $\{e_{ik}\}_{k=1}^K$ and $\{\bar{e}_{ik}\}_{k=1}^K$ be the evidence after fusion of $\{e_{ik}^v\}_{k=1}^K$ with the initial evidence $\{e_{ik}^o\}_{k=1}^K$, with and without filtering, respectively.*

Taking A-CBF as an example, the overall uncertainty score under the two scenarios is given by:

$$u_i = \frac{K}{\sum_{k=1}^{K} (e_{ik}^o + \alpha \cdot e_{ik}^v) + 1}, \bar{u}_i = \frac{K}{\sum_{k=1}^{K} (e_{ik}^o + e_{ik}^v) + 1}.$$

Since $\alpha \in (0, 1)$, filtering reduces the contribution of unreliable evidence e_{ik}^v in the denominator, resulting in $u_i > \bar{u}_i$. This demonstrates that evidence filtering increases the overall uncertainty score compared to fusion without filtering for unreliable views.

Similarly, for the BCF and ABF fusion strategies, filtering with $\alpha < 1$ consistently reduces the impact of unreliable evidence in the denominator, leading to the inequality $u_i > \bar{u}_i$. This indicates that evidence filtering amplifies the model's uncertainty for unreliable views, making it more cautious in its predictions and less influenced by noisy views. □