Divisiveness-Consistent Label Distribution Learning

Yunan Lu[1], Haitao Wu[2], Weiwei Li[3], Lei Yang[1], Xiuyi Jia[2]

[1] The Hong Kong Polytechnic University
[2] Nanjing University of Science and Technology
[3] Nanjing University of Aeronautics and Astronautics

ICML 2026

Motivation

Background: Label Distribution Learning

- Label Distribution Learning (LDL) is an effective learning paradigm for predicting entire label distributions, improving the trust-worthiness of predictions in risk-sensitive tasks.
- The objective of LDL is to obtain a simplex-valued regressor, which is trained on a set of instances annotated with label distributions.
  - Label distribution is a vector like probability distribution, where each element denotes the description degree of the label to the instance, such as $[0.2, 0.3, 0.5]$, $[0.1, 0.4, 0.5]$.
  - A simplex-valued regressor is a mapping from instance features to label distributions.

Motivation

Current Research Gap: Divisiveness

- Previous work neglects a critical characteristic underlying the label distribution: **Divisiveness**.
- **Divisiveness**: the propensity of label distribution to exhibit dissension between semantically opposing labels.
- Two examples of why divisiveness matters:
  - In the **movie rating prediction**, task where the ratings represent the crowd preference for a movie, a U-shaped rating distribution (where ratings are concentrated at both high and low extremes) signals intense polarization of crowd preference, which entails **greater commercial risk and opportunity** than a bell-shaped distribution (where ratings cluster near the median).
  - In the **urban governance**, policies characterized by U-shaped public opinion distributions, reflecting a high degree of polarization, tend to provoke **widespread social controversy and significant implementation challenges**.

Motivation

Current Research Gap: Divisiveness

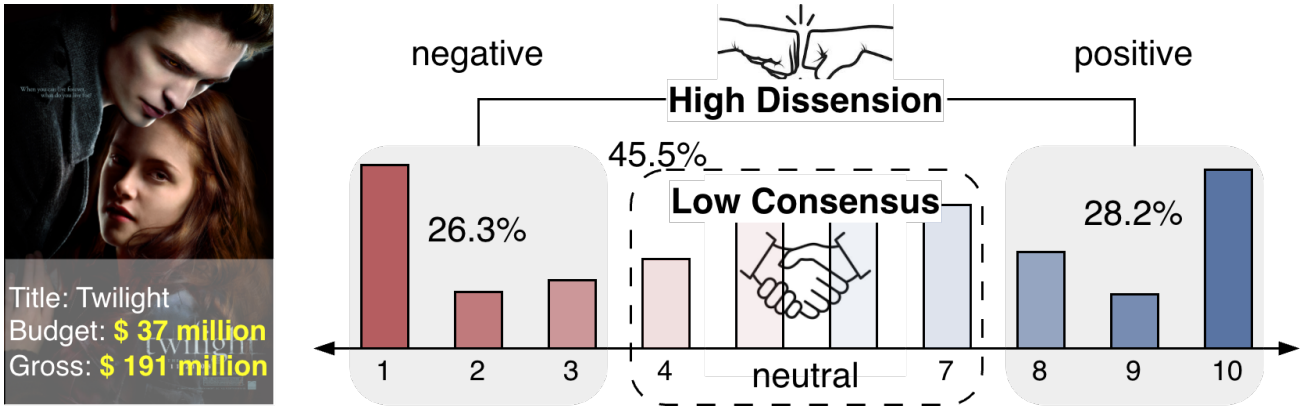

Figure 1: 26.3% opponents vs. 28.2% supporters ⇒ polarization (high risk).

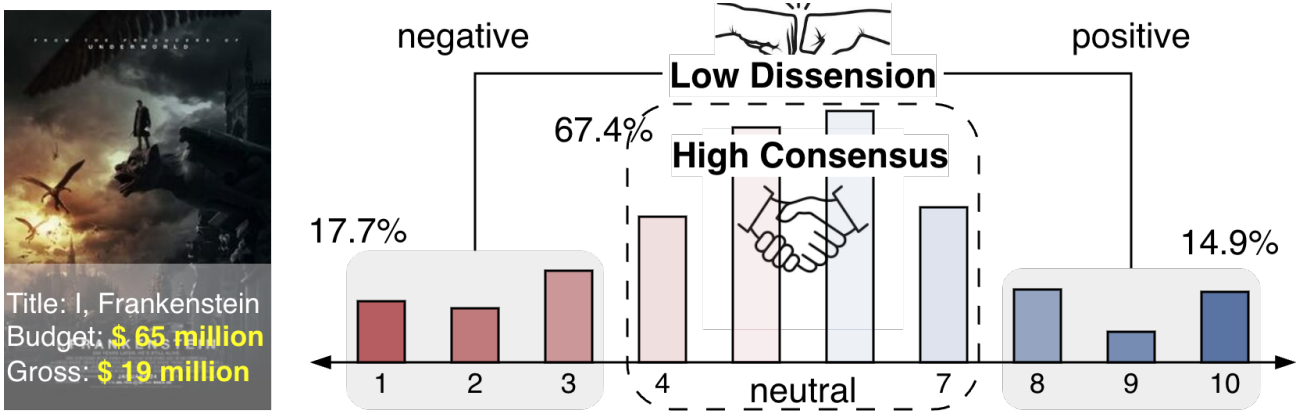

Figure 2: 17.7% opponents vs. 14.9% supporters ⇒ consensus (low risk).

Motivation and Background

Why KL or MSE Mismatch Divisiveness

- Kullback–Leibler (KL) divergence and Mean Squared Error (MSE), widely employed in current LDL methods, focus on the global distributional proximity. This objective is inherently inconsistent with the divisiveness.

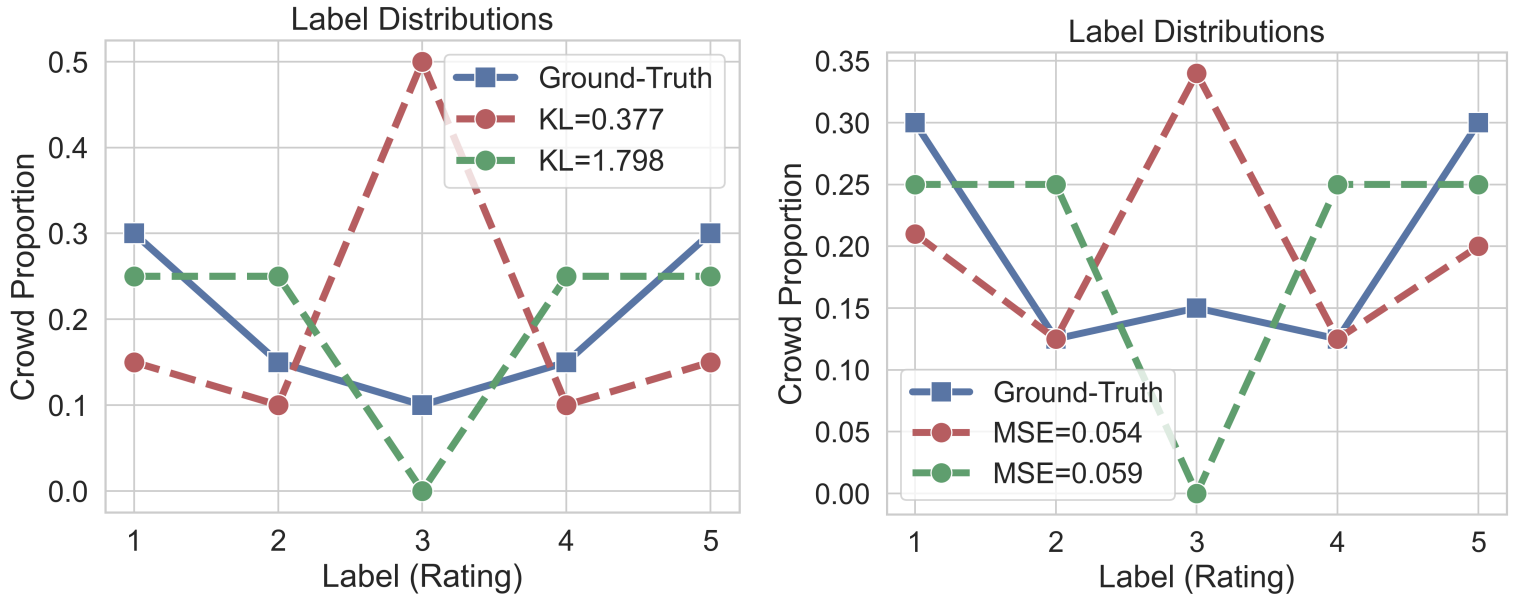

Figure 3: Two Examples of the metric inconsistency. Lower KL or MSE cannot denote lower divisiveness error.

Methodology

Measurement of Divisiveness

- Representative Case Analysis:

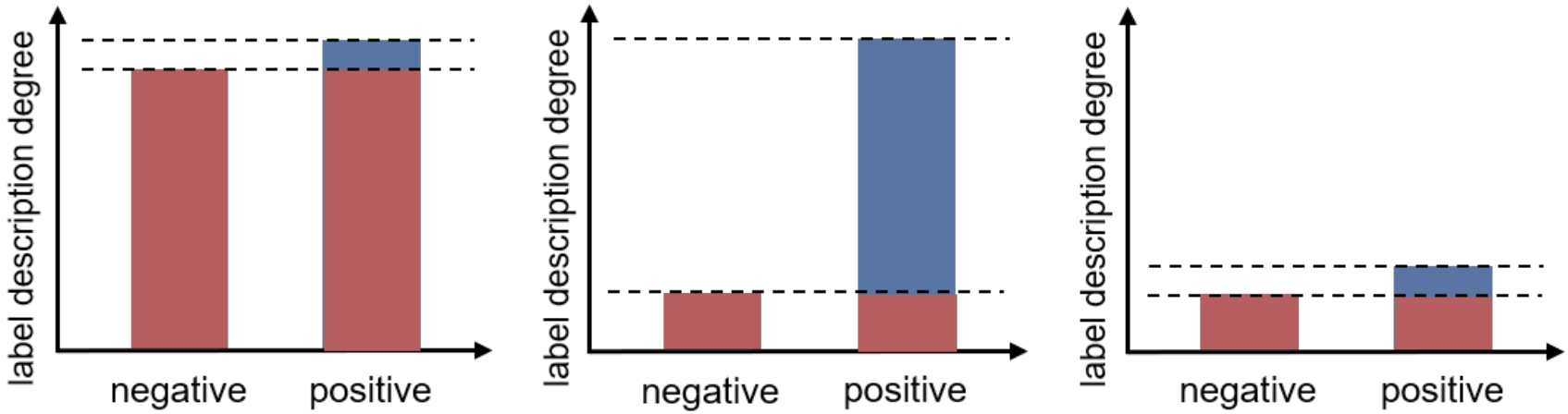

Figure 4: (left): High vs. High $\Rightarrow$ High Divisiveness; (mid): Low vs. High $\Rightarrow$ Low Divisiveness; (right): Low vs. Low $\Rightarrow$ Low Divisiveness;

- **Key Observation:** *Divisiveness between the opposing labels is positively related to the minimum of the description degree values of the two labels, i.e., the red portion.*

Methodology

Measurement of Divisiveness

- **Definition of Divisiveness**:

$$\psi(\boldsymbol{y}; \boldsymbol{\rho}, \boldsymbol{\eta}) = \min\{\langle \boldsymbol{\rho}, \boldsymbol{y} \rangle, \langle \boldsymbol{\eta}, \boldsymbol{y} \rangle\}, \tag{1}$$

  where each element in $\boldsymbol{\rho}$ denotes the degree to which each label belongs to the positive label group, each element in $\boldsymbol{\eta}$ denotes the degree to which each label belongs to the negative label group.
- **A Key Property**: Equation (1) preserves the *polarity monotonicity*.
- **Polarity Monotonicity**: If the value of label description degree shifts from positive or negative labels (where ρ or η is large) to labels that are more neutral (where ρ or η is smaller), the divisiveness of label distribution will decrease.
  - **Formal Definition of Polarity Monotonicity**: Let $h(\boldsymbol{y}; \boldsymbol{\rho}, \boldsymbol{\eta})$ denote a function for measuring the divisiveness within a label distribution $\boldsymbol{y}$ under positivity vector $\boldsymbol{\rho}$ and negativity vector $\boldsymbol{\eta}$. $h(\boldsymbol{y}; \boldsymbol{\rho}, \boldsymbol{\eta})$ is called to preserve the *polarity monotonicity*, if $h(\boldsymbol{y}; \boldsymbol{\rho}, \boldsymbol{\eta}) \leq h(\boldsymbol{y}^{\{i \to j\}}; \boldsymbol{\rho}, \boldsymbol{\eta})$ holds for any negatively directed redistribution $\boldsymbol{y}^{\{i \to j\}}$ and $\rho_i \leq \rho_j$, or for any positively directed redistribution $\boldsymbol{y}^{\{i \to j\}}$ and $\eta_i \leq \eta_j$.

Methodology

Learning with Divisiveness (1/3)

- **Why Distance Loss Cannot Preserve Divisiveness Error?**

$$
\begin{aligned}
0 \leq |\psi(\boldsymbol{y};\boldsymbol{\rho},\boldsymbol{\eta}) - \psi(\tilde{\boldsymbol{y}};\boldsymbol{\rho},\boldsymbol{\eta})| &\leq \xi_\infty \cdot \sqrt{2\mathcal{D}_{\mathrm{KL}}\left(\tilde{\boldsymbol{y}}\|\boldsymbol{y}\right)}, \\
0 \leq |\psi(\boldsymbol{y};\boldsymbol{\rho},\boldsymbol{\eta}) - \psi(\tilde{\boldsymbol{y}};\boldsymbol{\rho},\boldsymbol{\eta})| &\leq \xi_\infty \cdot \|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|_1 , \\
0 \leq |\psi(\boldsymbol{y};\boldsymbol{\rho},\boldsymbol{\eta}) - \psi(\tilde{\boldsymbol{y}};\boldsymbol{\rho},\boldsymbol{\eta})| &\leq \xi_2 \cdot \|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|_2,
\end{aligned}
\tag{2}
$$

where $\mathcal{D}_{\mathrm{KL}}\left(\tilde{\boldsymbol{y}}\|\boldsymbol{y}\right)$ is the KL divergence between $\boldsymbol{y}$ and $\tilde{\boldsymbol{y}}$; $\|\cdot\|_1$ and $\|\cdot\|_2$ are the L_1 and L_2 vector norm, respectively; $\xi_\infty = \max\{\|\boldsymbol{\rho}\|_\infty, \|\boldsymbol{\eta}\|_\infty\}$, and $\xi_2 = \max\{\|\boldsymbol{\rho}\|_2, \|\boldsymbol{\eta}\|_2\}$.

- **Preliminary Definition of Divisiveness Loss**:

$$
\mathcal{L}_\psi = \mathcal{E}\left(\psi(\tilde{\boldsymbol{y}};\boldsymbol{\rho},\boldsymbol{\eta}) - \psi(\boldsymbol{y};\boldsymbol{\rho},\boldsymbol{\eta})\right), \tag{3}
$$

where $\mathcal{E}(\cdot)$ is an error function such as absolute error $|\cdot|$ or squared error $(\cdot)^2$.

Methodology

Learning with Divisiveness (2/3)

- **Shortcomings of Preliminary Definition**: Equation (3) yields identical gradient directions for all semantically parallel labels. For example, $\nabla_{\tilde{\boldsymbol{y}}}|\psi(\tilde{\boldsymbol{y}};\boldsymbol{\rho},\boldsymbol{\eta}) - \psi(\boldsymbol{y};\boldsymbol{\rho},\boldsymbol{\eta})| \in \{\pm\boldsymbol{\rho},\pm\boldsymbol{\eta}\}$. However, the gradient directions for individual labels inherently differ, since increasing the description degree of one label necessarily entails decreasing the description degrees of one or more other labels to maintain the simplex constraint of label distributions. Although the KL or MSE can partially compensate for such gradient inconsistency, they may still hinder the convergence efficiency to some extent, or even lead to non-convergence.

Methodology

Learning with Divisiveness (3/3)

- **Final Divisiveness-Preserved Loss**: We propose a pairwise divisiveness loss function as a surrogate for divisiveness error:

$$\hat{\mathcal{L}}_{\psi} = \frac{1}{m(m-1)} \sum_{1 \le i \ne j \le m} \mathcal{E}(\psi(\tilde{\boldsymbol{y}}; \boldsymbol{\rho} \odot \boldsymbol{e}_i, \boldsymbol{\eta} \odot \boldsymbol{e}_j) - \psi_{ij}^{\star}), \tag{4}$$

  where $\mathcal{E}(\cdot)$ is a error function such as squared error and absolute error, $\boldsymbol{e}_i$ denotes the one-hot vector with the i-th entry being 1, $\psi_{ij}^{\star} = \psi(\boldsymbol{y}; \boldsymbol{\rho} \odot \boldsymbol{e}_i, \boldsymbol{\eta} \odot \boldsymbol{e}_j)$ denotes the pairwise divisiveness computed from the ground-truth label distribution.

- **Smooth Version of Divisiveness-Preserved Loss**: We employ the following function to approximate the divisiveness measure, since the divisiveness measure is not globally differentiable:

$$\begin{aligned} \psi(\tilde{\boldsymbol{y}}; \boldsymbol{\rho}, \boldsymbol{\eta}) \approx & \left(\exp\left(-k\langle \boldsymbol{\rho}, \tilde{\boldsymbol{y}} \rangle\right) + \exp\left(-k\langle \boldsymbol{\eta}, \tilde{\boldsymbol{y}} \rangle\right) \right)^{-1} \\ & \times \left(\langle \boldsymbol{\rho}, \tilde{\boldsymbol{y}} \rangle \exp\left(-k\langle \boldsymbol{\rho}, \tilde{\boldsymbol{y}} \rangle\right) + \langle \boldsymbol{\eta}, \tilde{\boldsymbol{y}} \rangle \exp\left(-k\langle \boldsymbol{\eta}, \tilde{\boldsymbol{y}} \rangle\right) \right). \end{aligned} \tag{5}$$

Experiments

Main Performance (1/2)

Algorithms	KLD ↓		Cosine ↑		Spear. ↑		DVSE ↓	
				RAF-ML				
AA-*k*NN	● .4073	±.014	● .8221	±.006	● .6041	±.013	● .1170	±.004
RG4LDL	● .6784	±.015	● .6788	±.007	● .2841	±.024	● .1118	±.004
LDLSF	● .8256	±.069	● .9102	±.005	● .7713	±.010	● .1036	±.003
DF-LDL	● .1957	±.011	● .9273	±.005	● .7817	±.010	● .0997	±.004
LCLR	● .1957	±.011	● .9277	±.005	● .7839	±.010	● .0990	±.003
SA-BFGS	● .1957	±.011	● .9277	±.005	● .7840	±.010	● .0986	±.003
LDLLC	● .1956	±.011	● .9276	±.005	● .7842	±.010	● .0985	±.003
DPA	● .1957	±.011	● .9276	±.005	● .7841	±.010	● .0985	±.003
LRR	● .1957	±.011	● .9276	±.005	● .7841	±.010	● .0984	±.003
δ-LDL	● .1929	±.011	● .9284	±.005	● .7847	±.010	● .0829	±.004
LDL-DVS	**.1873**	±.011	**.9300**	±.005	**.7863**	±.009	**.0764**	±.003
				Music				
DF-LDL	● .1719	±.026	● .8726	±.018	● .2688	±.080	● .0866	±.013
SA-BFGS	● .1912	±.030	● .8665	±.018	● .3030	±.072	● .0808	±.010
LCLR	● .1821	±.034	● .8738	±.019	● .3673	±.073	● .0804	±.012
DPA	● .1758	±.027	● .8746	±.017	● .3224	±.072	● .0791	±.011
LRR	● .1751	±.027	● .8750	±.017	● .3214	±.071	● .0785	±.010
LDLLC	● .1751	±.027	● .8751	±.017	● .3212	±.070	● .0776	±.011
LDLSF	● .6168	±.306	● .8718	±.018	● .3123	±.072	● .0774	±.009
AA-*k*NN	● .1258	±.021	● .9067	±.015	● .3989	±.081	● .0711	±.008
RG4LDL	● .1304	±.017	● .9023	±.011	● .3420	±.064	● .0674	±.008
δ-LDL	.1061	±.016	.9218	±.011	● .4942	±.058	.0573	±.008
LDL-DVS	**.1052**	±.015	**.9225**	±.010	**.4965**	±.052	**.0571**	±.007

Experiments

Main Performance (2/2)

Algorithms	KLD ↓		Cosine ↑		Spear. ↑		DVSE ↓	
			Painting					
δ-LDL	.5288	±.048	.7377	±.023	.3146	±.092	.1574	±.028
DPA	● .5999	±.067	● .7116	±.028	● .2780	±.069	.1529	±.019
LDLLC	● .5994	±.066	● .7120	±.027	● .2774	±.068	.1527	±.016
LDLSF	● 4.1107	±1.205	● .6565	±.033	● .2047	±.067	.1526	±.021
LRR	● .5988	±.064	● .7118	±.027	● .2782	±.069	.1501	±.017
RG4LDL	● .8089	±.126	● .6501	±.033	● .1870	±.070	.1478	±.018
LCLR	● .9065	±.215	● .6579	±.036	● .2454	±.066	.1471	±.019
SA-BFGS	● .8997	±.233	● .6596	±.036	● .2363	±.069	○ .1468	±.019
DF-LDL	● .6950	±.096	● .6772	±.032	● .2310	±.073	○ .1445	±.018
AA-kNN	● .7100	±.161	● .6987	±.027	● .2241	±.072	○ **.1416**	±.021
LDL-DVS	**.5262**	±.041	**.7389**	±.019	**.3188**	±.076	.1523	±.021
			emotion6					
AA-kNN	● .8328	±.072	● .6529	±.013	● .2257	±.031	● .1222	±.007
LDLSF	● .9999	±.123	● .6997	±.012	● .3125	±.030	● .1156	±.006
SA-BFGS	● .6204	±.025	● .6998	±.012	● .3257	±.028	● .1151	±.006
LCLR	● .6454	±.058	● .6870	±.030	● .2959	±.070	● .1151	±.008
DF-LDL	● .6162	±.025	● .7013	±.011	● .3308	±.028	● .1146	±.006
LDLLC	● .6129	±.023	● .7039	±.011	● .3325	±.030	● .1139	±.006
DPA	● .6241	±.023	● .6970	±.011	● .3138	±.030	● .1136	±.006
LRR	● .6111	±.025	● .7051	±.011	● .3323	±.031	● .1133	±.006
RG4LDL	● .6206	±.022	● .7007	±.009	● .3280	±.027	● .1105	±.006
δ-LDL	● .5636	±.024	● .7283	±.012	● .4051	±.027	● .1083	±.005
LDL-DVS	**.5624**	±.024	**.7288**	±.011	**.4061**	±.028	**.1074**	±.005

Conclusion

- **Principled Formalization of Divisiveness**.
  We formalize a divisiveness measure satisfying the axiomatic property of polarity monotonicity, which accurately captures the propensity for dissension between semantically opposing labels and provides a theoretical foundation for divisiveness-consistent learning.
- **Theoretical Analysis of Loss Inconsistency**. We theoretically analyze the inconsistency between conventional loss functions and divisiveness error, which reveals why optimizing distributional proximity alone cannot minimize divisiveness error.
- **Unbiased Surrogate of Divisiveness Error**. We propose a pairwise divisiveness loss as an unbiased surrogate of divisiveness error, which ensures stable convergence while effectively harmonizing distributional proximity and divisiveness error.