Addressing Imbalance in Multi-Label Classification Using Structured Hellinger Forests

Zachary A. Daniels, Dimitris N. Metaxas
zad7@cs.rutgers.edu, dnm@cs.rutgers.edu

Center for Computational Biomedicine Imaging and Modeling
Department of Computer Science
Rutgers, The State University of New Jersey

Problem Overview

- Binary multi-label classification involves finding a model that maps a set of input features $X \in \mathbb{R}^{1\times m}$ to more than one binary output label $Y \in \{0,1\}^{1\times n}$.
- Class imbalance is a common and challenging problem in multi-label classification which can be viewed from two perspectives: imbalance *between* labels and imbalance *within* labels.
- Our model, *Sparse Oblique Structured Hellinger Forests (SOSHF)*, extends structured forests [3] by explicitly addressing imbalanced data: incorporating cost-sensitivity into the clustering step and learning splits using a criterion based on the Hellinger distance [2].

Structured Forests

- Decision trees recursively partition a set of training instances based on some splitting criterion [4].
- Random forests ensemble multiple decision trees learned from different subsets of training instances and/or features [1].
- Structured forests are random forests used for structured prediction [3]. Structured forests learn splits at each node by transforming a set of multiple, structured labels to a single label and then optimizing a standard single label-based splitting criterion.

Proposed Work: Sparse Oblique Structured Hellinger Forests

Step 1: Cost-Sensitive Clustering

- Standard structured forests do not have a mechanism in place for dealing with imbalanced data.
- We incorporate cost-sensitivity into the transformation step of structured forests. To learn the transformation from multiple labels to a single label, we first weigh each label by a corresponding cost determined by it's inverse document frequency at the *global* (over all training data) and *local* (over the training data at a specific node) levels and then perform (weighted) k-means clustering.

$$\mathrm{IDF}(i) = \beta * \frac{\log\left(1+\frac{N_g}{n_{gi}}\right)}{\max_j \log\left(1+\frac{N_g}{n_{gj}}\right)} + (1-\beta) * \frac{\log\left(1+\frac{N_l}{n_{li}}\right)}{\max_j \log\left(1+\frac{N_l}{n_{lj}}\right)}$$

Step 2: The Sparse Hellinger Loss

- When using k-means clustering, we often get clusters of different sizes, trading the problem of imbalance over the original multi-label space for imbalance over the transformed, single-label space.
- Cieslak et al. showed that standard splitting criteria such as information gain are often ill-suited for imbalanced data and instead proposed using the Hellinger distance, a measure of separation between two probability distributions [2].
- The Hellinger distance for binary classification problems can be formulated using the true positive rate tpr and false positive rate fpr of some set of label assignments:

$$d_H(\mathrm{tpr},\mathrm{fpr}) = \sqrt{(\sqrt{\mathrm{tpr}}-\sqrt{\mathrm{fpr}})^2 + (\sqrt{1-\mathrm{tpr}}-\sqrt{1-\mathrm{fpr}})^2}$$

- *Note that if the* fpr *is higher than the* tpr*, we can flip the label assignments, and the distance doesn't change.*
- We want to learn oblique (linear, non-axis parallel) splits at each node. This requires learning a set of weights at each node that define a linear separating hyperplane.
- We propose a differentiable approximation of the Hellinger distance and use this approximation to define a loss function for learning the previously mentioned hyperplanes. We also incorporate a sparsity-inducing penalty on the weights of the hyperplane to prevent overfitting.
- *Please see the paper for additional technical details.*

Results

- We conduct experiments over ten datasets drawn from a wide range of domains with diverse characteristics and varying levels of imbalance (*see Table 1*).
- In general, Sparse Oblique Structured Hellinger Forests significantly outperform many classic and state-of-the-art multi-label algorithms with respect to the macro-F-Measure and the macro-AUC (*see Tables 2 and 3*).
- In particular, our proposed algorithm performs as well as COCOA [5] (the state-of-the-art for imbalanced multi-label classification) and cost-sensitive binary relevance forests (which require many more trees).

Dataset	Domain	|Instances|	|Features|	|Labels|	Type	LD	PDL	Min IR	Max IR	Mean IR
CAL500	Audio	502	68	124	Num	0.20	1.00	1.04	24.10	8.45
Emotions	Audio	593	72	6	Num	0.31	0.05	1.25	3.01	2.32
Medical	BioNLP	978	237	14	Nom	0.08	0.04	2.68	43.45	19.94
Enron	Text	1702	999	24	Nom	0.13	0.32	1.01	43.79	16.15
Scene	Images	2407	294	6	Num	0.18	0.01	3.52	5.61	4.66
Yeast	Bioinfo	2417	103	13	Num	0.32	0.08	1.33	12.58	4.25
Corel5k	Text	5000	499	44	Nom	0.05	0.21	3.46	49.00	29.40
RCV1: Subset1	Text	6000	1475	43	Num	0.06	0.10	3.34	49.42	25.53
RCV1: Subset2	Text	6000	1456	39	Num	0.06	0.08	3.22	47.78	26.37
TMC2007	Text/Sci	28596	278	15	Nom	0.14	0.02	1.45	34.26	13.58
Mediamill	Video	43907	120	29	Num	0.14	0.08	1.75	44.74	16.44

Table 1: Characteristics of the datasets used in our experiments.

	CAL	Emot	Med	Enron	Scene	Yeast	Corel	RCV1	RCV2	TMC*	Media*	Summary
svm-cost	0.261	0.632	0.765	0.354	0.633	0.473	0.216	0.356	0.345	0.584	0.352	(8.5)▲
svm-down	0.275	0.602	0.617	0.265	0.569	0.460	0.122	0.350	0.313	0.514	0.333	(12.5)▲
svm-adasyn	0.259	0.618	0.755	0.347	0.630	0.466	0.208	0.355	0.343	0.563	0.343	(9.9)▲
rf-cost	0.306	0.656	0.795	**0.411**	0.754	0.520	0.211	0.449	0.432	0.702	0.484	(3.2) △
rf-down	0.285	0.650	0.705	0.264	0.628	0.484	0.129	0.323	0.295	0.547	0.337	(11.5)▲
rf-adasyn	0.226	0.651	**0.801**	0.343	0.740	0.479	0.081	0.329	0.305	0.648	0.478	(7.9)▲
ml-knn	0.073	0.592	0.507	0.152	0.722	0.386	0.030	0.118	0.105	0.483	0.244	(16.9)▲
iblr	0.228	0.629	0.558	0.219	0.728	0.408	0.055	0.195	0.198	0.504	0.276	(14.5)▲
ecc	0.094	0.633	0.781	0.296	0.729	0.402	0.051	0.244	0.230	0.617	0.247	(12.5)▲
clr	0.083	0.593	0.768	0.290	0.633	0.408	0.048	0.233	0.233	0.610	0.265	(14.2)▲
rakel	0.191	0.613	0.766	0.307	0.692	0.428	0.087	0.309	0.298	0.623	0.374	(11.4)▲
homer	0.254	0.575	0.764	0.332	0.595	0.443	0.146	0.317	0.305	0.589	0.320	(12.1)▲
cocoa	0.228	0.660	0.777	0.389	0.743	0.479	0.200	0.390	0.376	0.656	0.454	(5.6)▲
sf	**0.308**	0.660	0.454	0.324	0.644	0.481	0.202	0.364	0.323	0.587	0.343	(8.3)▲
sf-lr	0.301	0.659	0.685	0.252	0.690	0.501	0.172	0.260	0.242	0.704	0.445	(9.4)▲
sf-lr-cs	0.301	0.653	0.727	0.247	0.694	0.505	0.198	0.319	0.301	0.719	0.405	(8.5)▲
sf-h	0.305	0.682	0.775	0.375	**0.763**	0.521	0.231	**0.476**	**0.453**	0.727	0.496	(2.5) △
sf-h-cs	0.306	**0.684**	0.792	0.379	0.758	**0.524**	**0.246**	0.472	0.452	**0.731**	**0.505**	(1.8) △

Table 2: Macro-f-measure of classifiers on benchmark datasets. Parentheses denote the mean rank. ▲ denotes SF-H-CS is statistically superior at $p = 0.05$ and △ denotes no significant difference according to the Friedman test with the Wilcoxon signed-rank posthoc test with Bonferroni correction. *COCOA run with ensemble size of 10 due to computational limitations

	CAL	Emot	Med	Enron	Scene	Yeast	Corel	RCV1	RCV2	TMC*	Media*	Summary
svm-cost	0.534	0.800	0.962	0.726	0.872	0.659	0.707	0.838	0.835	0.914	0.817	(10.5)▲
svm-down	0.533	0.774	0.967	0.705	0.864	0.642	0.689	0.897	0.889	0.905	0.798	(11.4)▲
svm-adasyn	0.532	0.793	0.959	0.721	0.868	0.653	0.700	0.837	0.834	0.907	0.813	(12.2)▲
rf-cost	0.552	0.838	0.963	0.779	0.939	0.703	0.728	0.899	0.892	0.932	0.835	(5.1)▲
rf-down	0.552	0.824	0.966	0.700	0.911	0.693	0.673	0.890	0.878	0.924	0.814	(9.0)▲
rf-adasyn	0.541	0.832	0.962	0.767	0.933	0.693	0.725	0.894	0.885	0.928	0.838	(7.2)▲
ml-knn	0.515	0.812	0.913	0.654	0.926	0.684	0.587	0.664	0.671	0.855	0.767	(14.8)▲
iblr	0.508	0.833	0.921	0.686	0.935	0.698	0.650	0.789	0.792	0.881	0.801	(12.4)▲
ecc	0.557	0.843	0.938	0.736	0.943	0.706	0.604	0.870	0.861	0.882	0.802	(8.5)▲
clr	0.562	0.794	0.965	0.759	0.896	0.651	0.739	0.898	0.891	0.905	0.805	(8.4)▲
rakel	0.529	0.798	0.900	0.679	0.894	0.640	0.550	0.738	0.726	0.850	0.736	(16.0)▲
homer	0.515	0.706	0.930	0.645	0.811	0.593	0.590	0.707	0.707	0.800	0.641	(17.0)▲
cocoa	0.560	0.839	0.969	**0.787**	0.941	**0.717**	0.732	0.911	0.905	0.928	0.842	(3.4) △
sf	0.559	0.835	0.916	0.736	0.901	0.643	0.741	0.880	0.860	0.901	0.755	(10.2)▲
sf-lr	0.549	0.835	0.960	0.655	0.909	0.683	0.685	0.797	0.785	0.938	0.828	(10.5)▲
sf-lr-cs	0.556	0.830	0.961	0.643	0.909	0.684	0.704	0.833	0.815	0.942	0.789	(10.6)▲
sf-h	0.569	0.848	0.976	0.771	0.943	0.708	0.736	**0.922**	**0.912**	0.945	**0.857**	(2.2) △
sf-h-cs	**0.572**	**0.848**	**0.977**	0.776	**0.943**	0.706	**0.746**	0.921	0.910	**0.946**	0.855	(1.6) △

Table 3: Macro-AUC of classifiers on benchmark datasets. Parentheses denote the mean rank. ▲ denotes SF-H-CS is statistically superior at $p = 0.05$ and △ denotes no significant difference according to the Friedman test with the Wilcoxon signed-rank posthoc test with Bonferroni correction. *COCOA run with ensemble size of 10 due to computational limitations

[1] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
[2] David A Cieslak, T Ryan Hoens, Nitesh V Chawla, and W Philip Kegelmeyer. Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*, 24(1):136–158, 2012.
[3] Piotr Dollár and C Lawrence Zitnick. Structured forests for fast edge detection. In *ICCV*, pages 1841–1848, 2013.
[4] J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.
[5] Min-Ling Zhang, Yu-Kun Li, and Xu-Ying Liu. Towards class-imbalance aware multi-label learning. In *IJCAI*, pages 4041–4047, 2015.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1433187.