

An Interactive Multi-Label Consensus Labeling Model for Multiple Labeler Judgments

Ashish Kulkarni^{1*}, Narasimha Raju Uppalapati², Pankaj Singh³, Ganesh Ramakrishnan⁴

¹Amazon, Bangalore, India. ²Samsung R&D Institute, Bangalore, India. ³Lokavida Technologies Pvt. Ltd.

⁴Department of Computer Science and Engineering, IIT Bombay, Mumbai, India.
{kulashish, raju.atl007, pr.pankajsingh}@gmail.com, ganesh@cse.iitb.ac.in

Abstract

Multi-label classification is crucial to several practical applications including document categorization, video tagging, targeted advertising *etc.* Training a multi-label classifier requires a large amount of labeled data which is often unavailable or scarce. Labeled data is then acquired by consulting multiple labelers—both human and machine. Inspired by ensemble methods, our premise is that labels inferred with high consensus among labelers, might be closer to the ground truth. We propose strategies based on interaction and active learning to obtain higher quality labels that potentially lead to greater consensus. We propose a novel formulation that aims to collectively optimize the cost of labeling, labeler reliability, label-label correlation and inter-labeler consensus. Evaluation on data labeled by multiple labelers (both human and machine) shows that our consensus output is closer to the ground truth when compared to the “majority” baseline. We present illustrative cases where it even improves over the existing ground truth. We also present active learning strategies to leverage our consensus model in interactive learning settings. Experiments on several real-world datasets (publicly available) demonstrate the efficacy of our approach in achieving promising classification results with fewer labeled data.

1 Introduction

Multi-label classification is a widely studied problem (Zhang and Zhou 2014) and naturally manifests in several real-world applications (Ueda and Saito 2002; Katakis, Tsoumakas, and Vlahavas 2008; Qi et al. 2007) that include text categorization, image or video tagging, information retrieval, recommendation systems, *etc.* Training of these models typically requires large amounts of labeled data which is often unavailable or scarce. Labeled data is then generated either by an expert oracle, by multiple humans or by a semi-automatic human-in-the-loop approach.

We consider this problem of labeling a large (possibly growing) multi-label data, beginning with the “cold start” setting, in which, there is little or no labeled data. We would like to label, with high confidence, all instances in the dataset and additionally, limit the labeling cost as much as possible. The problem poses several unique challenges.

*This work was conducted by the author at IIT Bombay prior to joining Amazon
Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Inter-label correlation: Certain labels, especially in a large label space (*E.g. lake-water*), might be mutually (semantically) correlated. This correlation may be exploited to minimize labeling cost (Xie et al. 2013).

Labeler reliability: We assume the availability of multiple labelers, not only humans, but also machine-learned multi-label classification models. These labelers might have complementary expertise, such that, the judgments by a labeler might be more reliable on a subset of labels. Can the labelers mutually benefit from their individual expertise (Raykar et al. 2010)?

Collective prediction: How do we aggregate the judgments from individual labelers, while accounting for inter-label correlation and labeler reliability to generate high confidence consensus label predictions for each instance? Table 1 shows an illustrative example from an image labeling task, where a simple majority-based consensus falls short of obtaining the expected labeled output.

Labeling cost: Can we use the consensus output to “actively” train classification models in order to minimize the overall labeling budget (Tong and Chang 2001)?

These challenges form the focus of our study. We argue that in a problem like this involving multiple (i) instances, (ii) labels and (iii) labelers, with varying degrees of noise or confusion, it is the “*consensus*” along these three dimensions that orients the labeling towards the *desired* ground truth.

2 Related Work

Active learning (Settles 2009; Tong and Koller 2002) is a well researched area. **Pool-based sampling** (Tong and Chang 2001), a popular active learning strategy, assumes the availability of a small set of labeled data and a large pool of unlabeled data. Two main measures used to evaluate informativeness for sampling instances are *query by uncertainty* (Tong and Chang 2001) and *query by committee* (Seung, Oppel, and Sompolinsky 1992).

Whereas the body of literature cited above targets single-label classification problems, active learning for multi-label classification has received some attention as well. Brinker (2006) uses a one-vs-all approach to multi-label classification and selects instances that **minimize the smallest SVM margin** amongst all the one-vs-all binary classifiers. Li, Wang, and Sung (2004) propose a mean max loss (MML) strategy that samples instances that lead to the maximum re-


Image	Judgments	Majority	Desired
	<ul style="list-style-type: none"> • L1: temple • L2: buildings castle nighttime plants reflection sky water window • L3: buildings plants reflection • L4: castle plants reflection sky • L5: buildings castle lake plants reflection water • L6: nighttime • L7: buildings plants reflection sky water • L8: buildings plants reflection water 	<ul style="list-style-type: none"> • buildings • plants • reflection • water 	<ul style="list-style-type: none"> • buildings • plants • castle • nighttime • reflection • sky • water

Table 1: Illustrative example: Majority labels fall short of the expected output

duction in expected loss. Yang et al. (2009) also propose a sampling strategy that aims to maximize the expected loss reduction and uses the SVM version space to measure loss reduction. Qi et al. (2008; 2009) proposed two-dimensional active learning algorithms for image classification, that select instance-label pairs to minimize the Bayesian classification error bound. Russakovsky, Li, and Fei-Fei (2015) recently presented a Markov Decision Process-based framework that seamlessly integrates computer vision models and human feedback for an object annotation task. Griffith et al. (2013) directly incorporate human feedback for policy shaping in a Bayesian Q-learning RL formalism.

In contrast to relying on an omniscient oracle for labeling, there is also work in the active learning literature that makes use of many imperfect labelers (Raykar et al. 2010; Yan et al. 2011; Dekel and Shamir 2009) and accounts for both labeler and model uncertainty (Sheng, Provost, and Ipeirotis 2008). Donmez and Carbonell (2008) propose a *proactive learning* method that jointly selects the optimal labeler and instance with a decision theoretic approach. Several *consensus-based* prediction combination algorithms (Gao et al. 2009; Li and Ding 2008) exist that combine multiple model predictions to counteract the effects of data quality and model bias. Recently, Xie et al. (2013) extended this to multi-label classification and proposed algorithms to consolidate the predictions of base models by maximizing model consensus and exploiting label-label correlations. There also exist label embedding approaches from the extreme classification literature (Yeh et al. 2017; Bhatia et al. 2015) that exploit inter-label correlation. While these approaches assume that the imperfect labeler’s knowledge is fixed, Fang et al. (2012) recently presented a self-taught active learning paradigm, where a crowd of imperfect labelers learn complementary knowledge from each other. However, they use instance-wise reliability of labelers to query only the most reliable labeler without any notion of consensus.

Our work differs from the aforementioned works in two ways - (1) we consider a more general case of multi-label active learning while generating consensus amongst multiple labelers—both human and machine; (2) the scope of our problem poses the interesting research challenge of generating high-quality labeled data, while accounting for labeler reliability, model uncertainty and label-label correlations. Whereas the literature attempts to address subsets of this problem, to the best of our knowledge, there is no work

that has looked at this problem in its entirety.

Our Contributions: Using a carefully articulated interplay of high consensus labeling of instances and reliable users, we propose a novel interactive multi-label consensus labeling model. We also discuss how this model could be used in active learning settings for multi-label classification with multiple labelers. We propose an *influence-based sampling* strategy that samples instances for labeling so as to maximize the expected consensus on unlabeled data. Additionally, a computationally cheaper, yet effective *uncertainty-based sampling* approach is proposed, so as to improve the consensus on the most confused instances in the unlabeled dataset.

3 Problem Definition

We are provided a set $X = \{\mathbf{x}_i\}_{i \in \{1 \dots n\}}$ of data instances, and a label set $L = \{y_j\}_{j \in \{1 \dots l\}}$. Each \mathbf{x}_i is a feature vector representing the corresponding instance. The task of multi-label classification is to assign to each instance \mathbf{x}_i , a subset of labels from L . Let the ground truth labelling for \mathbf{x}_i be represented by a binary label vector $\mathbf{z}_i = [z_i^1 \dots z_i^l]$, where $z_i^j \in \{0, 1\}$. If $\mathcal{Z} = \{0, 1\}^l$ denotes the set of all possible label combinations, then the multi-label classification problem can be expressed as a decision function $f : \mathbf{X} \rightarrow \mathcal{Z}$.

Learning the decision function f requires large amounts of labeled data and manually curating this labeled data incurs cost. Let D_L be a (partially) labeled dataset and D_U be a large unlabeled dataset. Further, suppose that we have a group of m independent labelers $M = \{w_k\}_{k \in \{1 \dots m\}}$, both human and machine, possibly noisy and untrained. We wish to expand the labeled dataset D_L by successively querying for labels of unlabeled data points from D_U . On querying a labeler $w_k \in M$, the response Y^k is an $n \times l$ binary matrix with $Y_{ij}^k \in \{0, 1\}$ denoting the class value of the i^{th} instance for the j^{th} label by the k^{th} labeler. Let g be a prediction combination function that takes as input multiple label prediction matrices Y^k and outputs a final prediction matrix U . Then our aim is to expand the labeled dataset D_L by labeling the unlabeled data points from D_U using an efficient sampling strategy, such that the cost of labeling is minimized, and simultaneously train machine labelers $w_k \in M$ which learn the prediction functions f^k . In fact, from the perspective of the machine labeler, this interactive labeling approach is very much related to active learning.

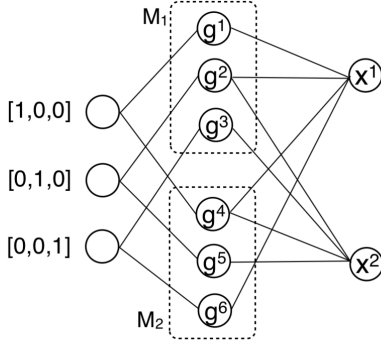


Figure 1: Bipartite Graph for MLCM-r

4 Our Approach

We assume the availability of a group of human/machine labelers with complementary knowledge sets such that, each might be more reliable on a subset of labels. While a majority-based voting scheme is popular in arriving at a combined prediction, we propose a consensus-based prediction combination model that also exploits the inter-label correlation and label-wise labeler reliability. Later, we also discuss how our consensus model could be leveraged in interactive settings and could be used to actively learn the machine models, thereby progressively lowering the human labeling cost. We present two active learning-based sampling strategies.

4.1 Consensus-based Prediction Combination

Our prediction combination function g leverages the multi-label consensus maximization for ranking (MLCM-r) approach proposed by (Xie et al. 2013) and we describe it briefly here. Consider a bipartite graph of n instance nodes $\mathbf{x}_1, \dots, \mathbf{x}_n$ and $v = m \times l$ group nodes, where a group node represents a labeler-label combination. Matrix A encodes the connection information in the bipartite graph, where an i^{th} instance is connected to a $(k-1) \times l + j$ -th group node (encoded as 1 in the matrix), if the k -th model predicts that the j -th label is relevant to the i -th instance (or $Y^k[i, j] = 1$). An instance node can be connected to more than one group nodes from a single model, representing the multi-label predictions of that model. Fig. 1 shows a sample graph for two instances, three labels and two labelers. A dashed rectangle surrounds group nodes which belong to the same model. Another layer of nodes, called label nodes, represent the label that a group node stands for. Connections between this layer of nodes and the group nodes is encoded in the matrix B . The matrices and notation used in MLCM-r are described in Table 2.

MLCM-r maximizes model consensus while exploiting label-label correlations. However, it assumes uniform labeler expertise and does not account for differential labeler reliability. In our consensus maximization formulation, we alter the original MLCM-r formulation in order to incorporate labeler reliability.

	Explanation
v	Number of groups ($= m \times l$), with index c
A	$n \times v$ matrix such that $a_{i,c}$ is the prediction of label $(c \bmod l)$ on instance \mathbf{x}_i by model $\lfloor c/l \rfloor$
B	$v \times l$ matrix of probability distributions on label nodes
U	$n \times l$ matrix such that $u_{i,j}$ is the probability that label j is relevant to \mathbf{x}_i
Q	$v \times l$ matrix such that $q_{c,j}$ is the probability of seeing label j given the label corresponding to group node g_c

Table 2: Notations for MLCM-r

4.2 Labeler Reliability

Labelers might have different expertise and therefore their labeling might be more reliable on certain labels than others. We propose a modified consensus model, referred to as MLCM^{ur}, that incorporates labeler reliability in the MLCM-r model. Let r_j^k (or r_c) be the reliability score of labeler M_k on label j (or group g_c where, $c = (k-1)l + j$). For each iteration $t' > 0$ we solve the following modified MLCM-r optimization problem, with r_c fixed

$$\min_{U, Q} \sum_{i=1}^n \sum_{c=1}^v r_c^{t'-1} \times a_{ic} \|\mathbf{u}_i^{t'} - \mathbf{q}_c^{t'}\|^2 + \alpha \sum_{c=1}^v \|\mathbf{q}_c^{t'} - \mathbf{b}_c\|^2 \quad (1)$$

s.t.

$$u_{ij}^{t'} \geq 0, \sum_{j=1}^l u_{ij}^{t'} = 1, i = 1, \dots, n$$

$$q_{cj}^{t'} \geq 0, \sum_{j=1}^l q_{cj}^{t'} = 1, c = 1, \dots, v$$

The superscript t' denotes the iteration index ($r_c^0 = 1$). $\mathbf{u}_i^{t'}$ and $\mathbf{q}_c^{t'}$ are obtained using equations (2) and (3).

The first term ensures that if an instance \mathbf{x}_i is linked to group g_c ($a_{ic} = 1$), then their conditional probability estimates must be close. The second term ensures that the probability distribution on the group nodes after consensus does not deviate much from its initial probability distribution. α is the penalty for constraint violation. \mathbf{u}_i and \mathbf{q}_c are probability vectors and therefore each of their components must be greater than or equal to 0 and their sum equals to 1.

The solution is obtained by block co-ordinate descent, where at every iteration t :

$$\mathbf{q}_c^t = \frac{\sum_{i=1}^n r_c^{t'-1} \times a_{ic} \mathbf{u}_i^{t-1} + \alpha \mathbf{b}_c}{\sum_{i=1}^n r_c^{t'-1} \times a_{ic} + \alpha} \quad (2)$$

$$\mathbf{u}_i^t = \frac{\sum_{c=1}^v r_c^{t'-1} \times a_{ic} \mathbf{q}_c^t}{\sum_{c=1}^v r_c^{t'-1} \times a_{ic}} \quad (3)$$

Upon convergence, the final probability distributions are given in the rows of U and the Q matrix captures the inter-label correlation probabilities. The labeler reliability is then updated in the following manner

$$r_c^{t'} \leftarrow r_c^{t'-1} + \gamma(\kappa_c^{t'} - r_c^{t'-1}) \quad (4)$$

where, κ_c is an agreement measure that captures the agreement between the k -th labeler and the consensus model on label j across all labeled instances and $\gamma < 1$ is a constant. The iterative procedure is repeated until convergence, $\|\kappa_c^{t'} - \kappa_c^{t'-1}\|^2 \leq \delta$, for some constant δ (we set it to 0.01). Both δ and γ are empirically tuned to maximize κ_c over currently labeled instances on a held-out dataset. As agreement measure, we use Cohen Kappa (Cohen 1968), which is a standard metric for inter-rater agreement.

5 Active Learning

Active learning aims to acquire labels for instances from an unlabeled pool, so as to train an underlying classifier, while incurring minimum labeling cost. It is an iterative procedure where at every successive iteration, it samples the most informative instance, **gathers its label** and trains the underlying classification model with the updated labeled instance pool. Here, we propose two strategies for sampling the most informative instance - (1) Uncertainty-based sampling and (2) Influence-based sampling. We then acquire its labels from the labelers and use our labeler reliability-weighted MLCM^{ur} model to arrive at the consensus labels for the instance.

5.1 Uncertainty-based Sampling

Let g_L be the consensus model obtained using m machine labelers, that are trained with the labeled dataset D_L . We also define an agreement function κ that measures agreement between label sets, where a label set could be an output $f_L^k(\mathbf{x}_i)$ of a k^{th} prediction function, output $g_L(\mathbf{x}_i)$ of the consensus model or the ground truth \mathbf{z}_i . Overall inter-labeler agreement for an instance \mathbf{x} is computed as

$$\kappa_{\mathbf{x}} = \frac{1}{m} \sum_{k=1}^m \kappa(f_L^k(\mathbf{x}), g_L(\mathbf{x})) \quad (5)$$

We then sample that instance from the unlabeled pool which has minimum overall agreement. That is, $\mathbf{x}^* = \min_{\mathbf{x} \in D_U} \kappa_{\mathbf{x}}$. Low inter-labeler agreement is an indication of confusion (or uncertainty) among machine labelers and therefore inclusion of this instance in the labeled pool (after human labeling) might lead to better trained machine labelers. The output of the consensus model, $g_L(\mathbf{x}) = \mathbf{u}_{\mathbf{x}}$, is a probability distribution across the labels. We obtain a binary label prediction from it, such that $\mathbf{u}_{\mathbf{x}}^j = 1$, if $\mathbf{u}_{\mathbf{x}}^j > \tau$, 0 otherwise, for some threshold $0 < \tau < 1$. The threshold is selected such that it maximizes the overall agreement on that instance, that is, $\tau = \arg \max_{0 < \tau < 1} \kappa_{\mathbf{x}}$.

5.2 Influence-based Sampling

The expected agreement on the unlabeled dataset D_U , for a consensus model obtained from D_L , is given by

$$\sigma_L = \frac{1}{|D_U|} \sum_{\mathbf{x} \in D_U} \sum_{\mathbf{z} \in \mathcal{Z}} \kappa(g_L(\mathbf{x}), \mathbf{z}) P(\mathbf{z}|\mathbf{x}) \quad (6)$$

After sampling an instance $\mathbf{x}^* \in D_U$, let $D_{L'} = D_L \cup \mathbf{x}^*$ be the new labeled dataset and the expected agreement on D_U

based on $D_{L'}$ is given by

$$\sigma_{L'} = \frac{1}{|D_U|} \sum_{\mathbf{x} \in D_U} \sum_{\mathbf{z} \in \mathcal{Z}} \kappa(g_{L'}(\mathbf{x}), \mathbf{z}) P(\mathbf{z}|\mathbf{x}) \quad (7)$$

We wish to sample \mathbf{x}^* that offers maximum improvement in the expected agreement.

$$\mathbf{x}^* = \arg \max_{\mathbf{x} \in D_U} \sum_{\mathbf{z} \in \mathcal{Z}} (\kappa(g_L(\mathbf{x}), \mathbf{z}) - \kappa(g_{L'}(\mathbf{x}), \mathbf{z})) P(\mathbf{z}|\mathbf{x})$$

Assuming that all instances in $D_U \setminus \{\mathbf{x}^*\}$ have an equal impact on the learners (Yang et al. 2009), we have

$$\mathbf{x}^* = \arg \max_{\mathbf{x} \in D_U} \sum_{\mathbf{z} \in \mathcal{Z}} (\kappa(g_L(\mathbf{x}), \mathbf{z}) - \kappa(g_{L'}(\mathbf{x}), \mathbf{z})) P(\mathbf{z}|\mathbf{x}) \quad (8)$$

The solution to the above equation requires us to estimate the improvement in expected agreement and estimate the conditional probability $P(\mathbf{z}|\mathbf{x})$. We will discuss these below.

Improvement in Expected Agreement (Dong et al. 2015) proposed an online consensus model that achieves on-line updates to the MLCM model by reformulating the iterative update equations as closed form solutions.

$$Q_A^* = (I - D_\lambda S_A)^{-1} D_{1-\lambda} B \quad (9)$$

$$U_{Q_A^*} = D_n^{-1} A Q_A^* \quad (10)$$

where,

$$D_v = \text{diag}\{\sum_{i=1}^n a_{ic}\}_{v \times v}, D_n = \text{diag}\{\sum_{c=1}^v a_{ic}\}_{n \times n}, K_v = \text{diag}\{\sum_{j=1}^l b_{cj}\}_{v \times v}, D_\lambda = (D_v + \alpha K_v)^{-1} D_v, D_{1-\lambda} = (D_v + \alpha K_v)^{-1} (\alpha K_v), S_A = D_v^{-1} A' D_n^{-1} A.$$

Consensus prediction for a new instance is then:

$$g_L(\mathbf{x}) = \frac{A_L(\mathbf{x}) Q_A^*}{1^T A_L(\mathbf{x})} \text{ and } g_{L'}(\mathbf{x}) = \frac{A_L(\mathbf{x}) Q_{A_x}^*}{1^T A_L(\mathbf{x})} \quad (11)$$

where, $A_L(\mathbf{x}) = [f_L^1(\mathbf{x}) f_L^2(\mathbf{x}) \cdots f_L^l(\mathbf{x})]$ is a $1 \times v$ matrix represents output of all prediction functions f_L and $\tilde{A}_x = \begin{bmatrix} A \\ A_L(\mathbf{x}) \end{bmatrix}$. $Q_{A_x}^*$ can be efficiently computed from Q_A^* by using the method proposed by Dong *et al.* This makes it possible to efficiently compute $g_{L'}(\mathbf{x})$ and thereby the agreement improvement $\kappa(g_L(\mathbf{x}), \mathbf{z}) - \kappa(g_{L'}(\mathbf{x}), \mathbf{z})$.

Estimate Conditional Probability Estimating $P(\mathbf{z}|\mathbf{x})$ for all possible $\mathbf{z} \in \mathcal{Z}$ is intractable due to the exponential search space ($\mathcal{Z} = \{0, 1\}^l$). This is particularly harder in an active learning setting due to limited training data. We therefore relax the search space by considering a subset $\mathcal{Z}_{\mathbf{x}} \subset \mathcal{Z}$ that represents the most possible label combinations for the instance \mathbf{x} . We expect the agreement to have maximum improvement on this subset $\mathcal{Z}_{\mathbf{x}}$, as the correct label combination might most likely belong to this subset. Next, we describe our relaxation approach to arrive at $\mathcal{Z}_{\mathbf{x}}$.

Assuming the labels to be independent, we have, $P(\mathbf{z}|\mathbf{x}) = \prod_j P(z^j|\mathbf{x})$. We model $P(z^j|\mathbf{x})$ as $h(\langle \mathbf{w}^j, \mathbf{x}'^j \rangle + w_0^j)$, where, $\mathbf{x}'^j \in \mathbb{R}^m$ is a feature vector corresponding to \mathbf{x} , comprising m labelers' output for the j -th label, w_0^j and \mathbf{w}^j are model parameters trained on the labeled data D_L . In

order to learn these weights, we train l logistic regression classifiers, using as features the m labelers' outputs in the labeled data. Next, we set z^j as given below.

$$z^j = \begin{cases} 0, & \text{if } P(z^j|\mathbf{x}) < 0.5 - \delta \\ 1, & \text{if } P(z^j|\mathbf{x}) \geq 0.5 + \delta \\ \{0, 1\}, & \text{if } 0.5 - \delta \leq P(z^j|\mathbf{x}) < 0.5 + \delta \end{cases}$$

It is this subset $\mathcal{Z}_{\mathbf{x}}$ of label combinations on which we compute the expected agreement. That is, we relax the Equation (8) to:

$$\mathbf{x}^* = \arg \max_{\mathbf{x} \in D_U} \sum_{\mathbf{z} \in \mathcal{Z}_{\mathbf{x}}} (\kappa(g_L(\mathbf{x}), \mathbf{z}) - \kappa(g_{L'}(\mathbf{x}), \mathbf{z})) P(\mathbf{z}|\mathbf{x})$$

6 Evaluation

We describe the experimental setup and present an ablation study of various components of our approach.

6.1 Experimental Setup

We evaluated our approach on multiple publicly available¹ datasets for multi-label classification (Table 3).

Dataset	#Instances	#Features	#Labels
Scene [†]	500	128	33
Flags [†]	194	19	7
Medical*	978	1,449	45
Enron*	1,672	1,001	53
Slashdot*	3,782	1,101	22
Corel5k*	5,000	499	374
Mediamill*	43,907	120	101

*Simulated labelers

[†]Human labelers

Table 3: Datasets

Human labeler judgments: Multiple human labelers' judgments are obtained by setting up labeling tasks on crowdflower² for two of the datasets *viz.* Flags and Scene (Chua et al. July 8 10 2009). We received human judgments from 8 labelers. For other datasets, we simulated labelers using an approach similar to that used by (Li, Jiang, and Zhou 2015) and describe it briefly next.

Simulation of human labelers with varied reliability: We simulated 6 human labelers with different expertise (reliability). For each dataset and for each label j , a logistic regression model is trained using all the instances and features in the dataset. The probability output of the model on the dataset is then used to obtain 6 clusters using k-means clustering. Each simulated human labeler $w_m, m \in \{1, \dots, 6\}$, is assumed to be an expert on the m -th cluster and provides the ground truth label for instances in that cluster. For the remaining data in the other clusters, it provides the ground truth label with probability 0.75 (or flips the ground truth label with probability 0.25).

Modeling machine labelers: We use a one-vs-all SVM as our choice of machine model. In order to simulate multiple machine models (we simulate 6 machine models) with

differential behavior on label subsets, we use an approach similar to the one for simulating human labelers.

6.2 Evaluation Measure

We use micro-averaged F_1 , a standard measure for evaluating multi-label classifiers (Yang et al. 2009).

$$F_1 = F_1^{micro} = \frac{2 \sum_{j=1}^l \sum_{i=1}^n y_i^j z_i^j}{\sum_{j=1}^l \sum_{i=1}^n y_i^j + \sum_{j=1}^l \sum_{i=1}^n z_i^j} \quad (12)$$

Dataset	Judgments		Majority		MLCM-r		MLCM ^{ur}	
	κ	F_1	κ	F_1	κ	F_1	κ	F_1
Flags	.6 - .8	.83-.9	.874	.934	.874	.934	.873	.933
Scene	.32 - .6	.35 - .65	.637	.675	.645	.682	.641	.679
Scene*	.34 - .79	.37 - .81	.824	.846	.851	.869	.894	.907

*Corrected ground truth

Table 4: How good is the consensus output

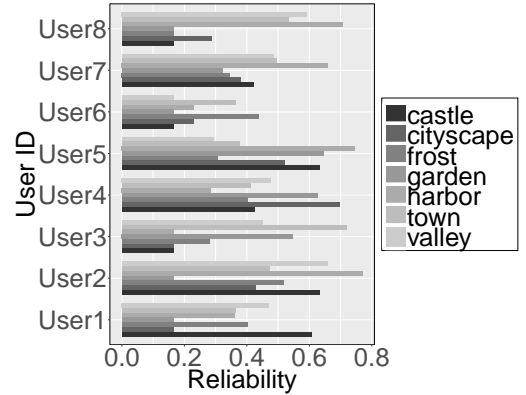


Figure 2: User reliability scores - Scene Dataset

6.3 How Good is Our Consensus Model?

We obtained consensus labels over the human judgments collected on the Flags and Scene datasets. The human judgments were subjected to three consensus models *viz.* (1) baseline majority-based consensus; (2) MLCM-r (without labeler reliability); (3) our MLCM^{ur}. Table 4 shows these results in comparison to the min-max κ (and F_1) obtained from labeler judgments. The consensus output is clearly better than any of the individual labeler judgments, thereby establishing the merit of consensus in general. Further, both MLCM-r and MLCM^{ur}, by virtue of modeling label-label correlations (Refer to Fig. 4), seem to outperform the baseline majority-based model. Although, both MLCM-r and MLCM^{ur} seem to show comparable performance, a closer inspection of the *Scene* ground truth revealed several discrepancies. We therefore, manually corrected this ground truth³ (*Scene** in the table) and evaluated the model outputs against this. Our MLCM^{ur} seems to benefit from the labeler reliability-based weighting and the consensus labels

¹<https://goo.gl/E49amv>

²<https://www.crowdfunder.com/>

³Shared at <https://goo.gl/ZG3gff>

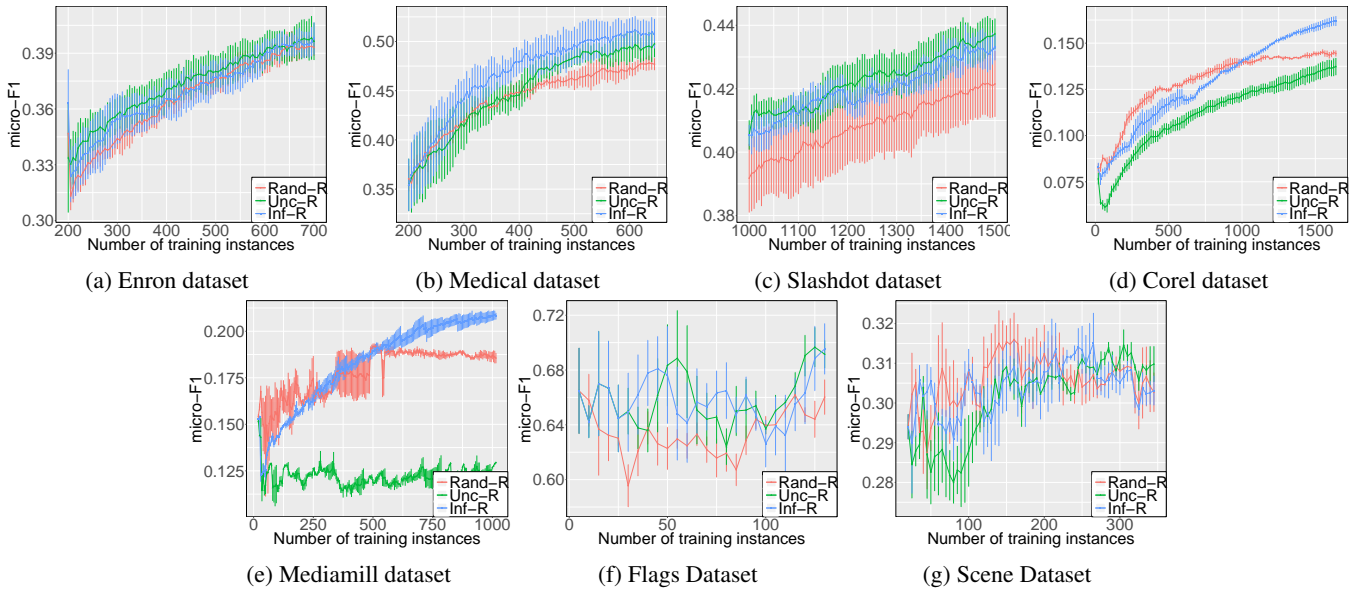


Figure 3: Average Micro- F_1 (with standard error) for different learning strategies on different datasets

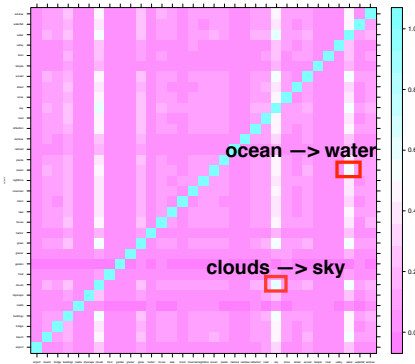


Figure 4: Label-label correlation in Scene dataset

indeed seem much closer to the ground truth. Further improvements in recall (8%) could be achieved by giving feedback to the labelers in cases where highly reliable labelers contradict with each other and prompting them to reconsider their judgments.


Image	Majority	MLCM-r	MLCM ^{ur}
	<ul style="list-style-type: none"> buildings plants reflection water 	<ul style="list-style-type: none"> buildings nighttime reflection sky water 	<ul style="list-style-type: none"> buildings castle nighttime reflection sky water

Table 5: Model comparison: Illustrative example

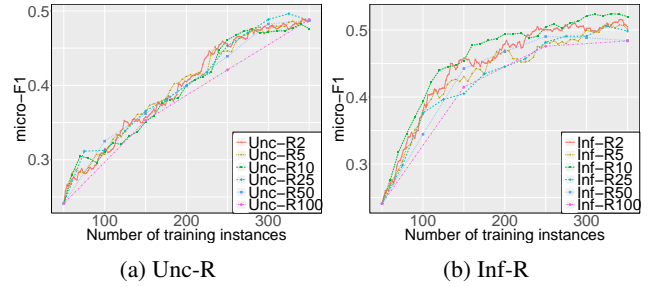


Figure 5: Effect of varying active learning batch size

6.4 Effect of Labeler Reliability

Fig. 2 shows the label-wise reliability scores (for a subset of labels) for each labeler computed by our MLCM^{ur} model on the Scene dataset. As can be seen, labeler reliability indeed seems to vary across labels. A few high reliability labelers for a label seem to influence the consensus output of the model (Refer to Table 5).

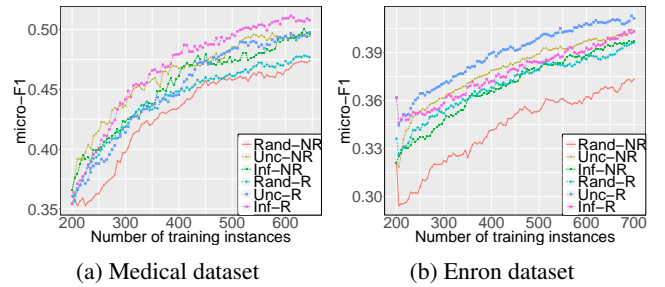


Figure 6: Effect of labeler reliability

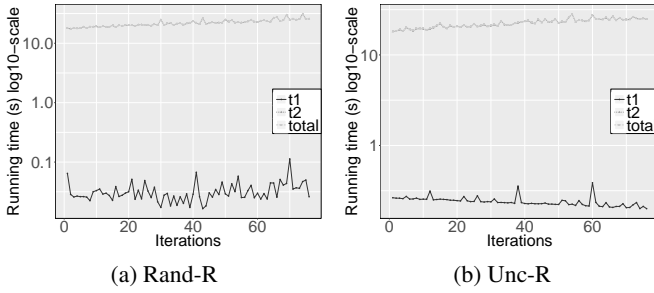


Figure 7: Effect of size of unlabeled instance space (iteration) on running time

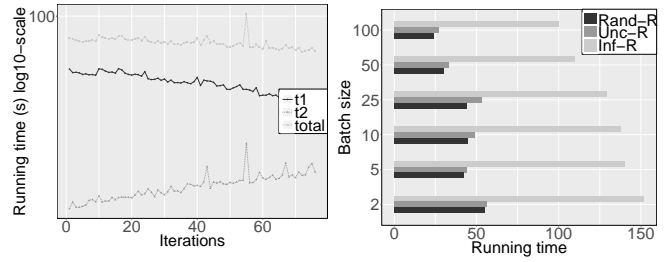


Figure 8: Effect of Batch size on running time (min.)

6.5 Evaluation of Active Learning

We leverage our consensus model in active learning settings using multiple human and machine labelers. The consensus output over multiple human judgments on a small subset of data is used to bootstrap cold-start scenarios. Using this labeled data to train the machine models, subsequent instances are sampled for labeling either at random (Rand-R) (for passive learning) or actively using (i) uncertainty-based sampling (Unc-R) or (ii) influence-based sampling (Inf-R).

Comparison of Passive and Active Learning Starting with an initial set of 20 labeled instances, we perform 100 iterations of learning, sampling a batch of 5 (15 for Corel) instances at every iteration. At every iteration, we evaluate the model on an independent test set (30% of dataset). The consensus output over the label prediction from multiple machine models is used for evaluation. We report average results (with standard error) over three runs with random initialization of the training data (Refer to Fig. 3). We see that the model does benefit from active learning and achieves higher accuracy with fewer number of training instances than those required for passive learning. As expected, Inf-R outperforms Unc-R in most cases due to its sampling of instances that lead to maximum improvement in expected agreement. The performance of Unc-R depends on the accuracy of trained models and might suffer if the models are not well trained. With less initial training data, the three policies seem to show a comparable performance. Practical applications could take benefit of faster Rand-R (Refer to Fig. 8) in the initial iterations and switch to Unc-R or Inf-R in the later iterations when the models are well trained.

Effect of Batch Size on Active Learning We also varied the sampling batch size and evaluated its impact on the performance of the active learner. Starting from 50 labeled instances, we ran active learning iterations until the labeled data size reached 350. The sampling batch size was set to 2, 5, 10, 25, 50 and 100. We observe (Refer to Fig. 5) that a smaller batch size generally results in better performance. Smaller batch sizes offer more opportunities to evaluate the unlabeled data on continuously improving models, thereby sampling the most informative instances. However, smaller batch sizes also result in higher run times. We study this trade-off in the next section.

Comparison of MLCM-r and MLCM^{ur} in Active Learning We compare the no reliability (NR) MLCM-r model

with our reliability (R) weighted MLCM^{ur} consensus model. Combined with the three sampling policies, this leads to six settings presented in Fig. 6. Reliability weighted model incorporating labeler reliability indeed leads to a better consensus and more accurate labeling.

Run-Time Analysis We profiled the impact of batch size on running time for each of three sampling strategies. All experiments were executed on a 38 core, 2.2 GHz server with 64 GB RAM. We started with 50 labeled instances of the medical dataset and then for different batch sizes, we ran the learning iterations until we had 350 labeled instances. We observe that as the batch size increases, running time decreases for each strategy (Refer to Fig. 8). As expected, Inf-R takes significantly more time as it involves evaluation of the impact of labeling of each unlabeled instance on the expected agreement.

In addition, we study the run-time of each iteration for a fixed sampling batch size. Starting with 1000 labeled instances, we run 75 iterations of learning, sampling five instances in each iteration. We report time t_1 taken for searching the unlabeled instance space and time t_2 for updating the consensus model for the selected batch of instances (Refer to Fig. 7). As expected, t_1 decreases as the number of unlabeled instances progressively reduce with each learning iteration. In the case of Inf-R, t_1 dominates the total running time and hence the total running time for an iteration progressively decreases. In case of Rand-R and Unc-R, t_1 is much lower than t_2 (t_2 overlaps with total time) and we therefore see a gradual rise in their total running time.

7 Conclusion

We considered the problem of multi-label classification in the absence of training data and oracles, and proposed an approach involving an interplay of consensus maximization and active learning. We may envisage a system that evolves in three phases (i) bootstrapping, using consensus of human labelers for initial unlabeled data (ii) active learning of machine labelers (iii) transitioning into an automatic system with minimal dependence on human labelers.

8 Acknowledgments

This work was funded by the Industrial Research Consultancy Center (IRCC), Indian Institute of Technology Bombay, India, through the Intranet Search Project, as well as

the IBM faculty award grant. We also acknowledge Simoni Shah and Aditya Kumar Akash for insightful discussions.

References

- Bhatia, K.; Jain, H.; Kar, P.; Varma, M.; and Jain, P. 2015. Sparse local embeddings for extreme multi-label classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems, NIPS'15*, 730–738. Cambridge, MA, USA: MIT Press.
- Brinker, K. 2006. On active learning in multi-label classification. In *From Data and Information Analysis to Knowledge Engineering*. Springer. 206–213.
- Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y.-T. July 8-10, 2009. Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*.
- Cohen, J. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* 70(4):213.
- Dekel, O., and Shamir, O. 2009. Good learners for evil teachers. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*, 233–240. New York, NY, USA: ACM.
- Dong, B.; Xie, S.; Gao, J.; Fan, W.; and Yu, P. S. 2015. On-linecm: Real-time consensus classification with missing values. *Proceedings of the 2015 SIAM International Conference on Data Mining* (685-693).
- Donmez, P., and Carbonell, J. G. 2008. Proactive learning: Cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08*, 619–628. New York, NY, USA: ACM.
- Fang, M.; Zhu, X.; Li, B.; Ding, W.; and Wu, X. 2012. Self-taught active learning from crowds. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, 858–863. IEEE.
- Gao, J.; Liang, F.; Fan, W.; Sun, Y.; and Han, J. 2009. Graph-based consensus maximization among multiple supervised and unsupervised models. In *Advances in Neural Information Processing Systems*, 585–593.
- Griffith, S.; Subramanian, K.; Scholz, J.; Isbell, C.; and Thomaz, A. L. 2013. Policy shaping: Integrating human feedback with reinforcement learning. In *Advances in Neural Information Processing Systems*, 2625–2633.
- Katakis, I.; Tsoumakas, G.; and Vlahavas, I. 2008. Multilabel text classification for automated tag suggestion. *ECML PKDD discovery challenge* 75.
- Li, T., and Ding, C. 2008. Weighted consensus clustering. *Proceedings of the 2008 SIAM International Conference on Data Mining* 1(2).
- Li, S.; Jiang, Y.; and Zhou, Z. 2015. Multi-label active learning from crowds. *CoRR* abs/1508.00722.
- Li, X.; Wang, L.; and Sung, E. 2004. Multilabel svm active learning for image classification. In *Image Processing, 2004. ICIP'04. 2004 International Conference on*, volume 4, 2207–2210. IEEE.
- Qi, G.-J.; Hua, X.-S.; Rui, Y.; Tang, J.; Mei, T.; and Zhang, H.-J. 2007. Correlative multi-label video annotation. In *Proceedings of the 15th international conference on Multimedia*, 17–26. ACM.
- Qi, G.-J.; Hua, X.-S.; Rui, Y.; Tang, J.; and Zhang, H.-J. 2008. Two-dimensional active learning for image classification. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8. IEEE.
- Qi, G.-J.; Hua, X.-S.; Rui, Y.; Tang, J.; and Zhang, H.-J. 2009. Two-dimensional multilabel active learning with an efficient online adaptation model for image classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31(10):1880–1897.
- Raykar, V. C.; Yu, S.; Zhao, L. H.; Valadez, G. H.; Florin, C.; Bogoni, L.; and Moy, L. 2010. Learning from crowds. *The Journal of Machine Learning Research* 11:1297–1322.
- Russakovsky, O.; Li, L.-J.; and Fei-Fei, L. 2015. Best of both worlds: human-machine collaboration for object annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2121–2131.
- Settles, B. 2009. Active learning literature survey. *Computer Sciences Technical Report 1648, University of Wisconsin, Madison* 52(55-66):11.
- Seung, H. S.; Oppor, M.; and Sompolinsky, H. 1992. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92*, 287–294. New York, NY, USA: ACM.
- Sheng, V. S.; Provost, F.; and Ipeirotis, P. G. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08*, 614–622. New York, NY, USA: ACM.
- Tong, S., and Chang, E. 2001. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, 107–118. ACM.
- Tong, S., and Koller, D. 2002. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research* 2:45–66.
- Ueda, N., and Saito, K. 2002. Parametric mixture models for multi-labeled text. In *Advances in neural information processing systems*, 721–728.
- Xie, S.; Kong, X.; Gao, J.; Fan, W.; and Yu, P. S. 2013. Multilabel consensus classification. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, 1241–1246. IEEE.
- Yan, Y.; Fung, G. M.; Rosales, R.; and Dy, J. G. 2011. Active learning from crowds. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 1161–1168.
- Yang, B.; Sun, J.-T.; Wang, T.; and Chen, Z. 2009. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 917–926. ACM.
- Yeh, C.-K.; Wu, W.-C.; Ko, W.-J.; and Wang, Y.-C. F. 2017. Learning deep latent spaces for multi-label classification.
- Zhang, M.-L., and Zhou, Z.-H. 2014. A review on multi-label learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on* 26(8):1819–1837.