Scalable Optimization of Multivariate Performance Measures in Multi-instance Multi-label Learning

Apoorv Aggarwal\textsuperscript{*} and Sandip Ghoshal\textsuperscript{*} and Ankith M. S. and Suhit Sinha and Ganesh Ramakrishnan
Indian Institute of Technology Bombay
\{apoorvaggarwal21, sandip.iem, suhit.iitb, msankith\}@gmail.com, ganesh@cse.iitb.ac.in

Purushottam Kar
Indian Institute of Technology Kanpur
purushot@cse.iitk.ac.in

Prateek Jain
Microsoft Research
prajain@microsoft.com

Abstract
The problem of multi-instance multi-label learning (MIML) requires a bag of instances to be assigned a set of labels most relevant to the bag as a whole. The problem finds numerous applications in machine learning, computer vision, and natural language processing settings where only partial or distant supervision is available. We present a novel method for optimizing multivariate performance measures in the MIML setting. Our approach MIML\textsuperscript{perf} uses a novel plug-in technique and offers a seamless way to optimize a vast variety of performance measures such as macro and micro-F measure, average precision, which are performance measures of choice in multi-label learning domains. MIML\textsuperscript{perf} offers two key benefits over the state of the art. Firstly, across a diverse range of benchmark tasks, ranging from relation extraction to text categorization and scene classification, MIML\textsuperscript{perf} offers superior performance as compared to state of the art methods designed specifically for these tasks. Secondly, MIML\textsuperscript{perf} operates with significantly reduced running times as compared to other methods, often by an order of magnitude or more.

1 Introduction
The paucity of labeled data has fueled much interest in learning paradigms with partial or distant supervision. The task of Multi-instance Multi-label (MIML) learning is one such paradigm that has received much attention in areas such as relation extraction (RE), image classification etc. The problem requires a bag of data points to be assigned a set of labels most relevant to the bag as a whole. To take an example, in the RE setting, we are interested in identifying relations held by entities being discussed in a body of text. The problem is pretty straightforward in the presence of sentence level annotations \textit{i.e.}, indications of which sentences discuss specific relations. However, this requirement is a prohibitive one, especially when the number of relations runs into the hundreds or thousands. Consequently, the problem is often posed as that of learning under distant supervision (Haffari, Nagesh, and Ramakrishnan 2015; Hoffmann et al. 2011; Surdeanu et al. 2012; Zeng et al. 2015) wherein one is pro-vided only with relational facts in a database along with a non-annotated corpus that overall, supports most of the facts.

The general MIML problem has been explored through reductions to single instance or single label learning in the past (Zhou and Zhan 2006; Zhan and Zhou 2008). The RE problem has enjoyed more focused attention and several diverse approaches have been proposed for this problem such as structural risk minimization Hoffmann et. al. (2011), neural networks (Zeng et al. 2015) and graphical models (Surdeanu et al. 2012).

Existing approaches for the MIML problem suffer from two main drawbacks 1) Although performance measures such as F-measure and average precision are standard for evaluation, the algorithms do not seek to optimize them directly, instead choosing to adopt heuristics that encourage good performance. An exception is the work of Haffari et al. (2015) who attempt to directly optimize the F-micro measure. 2) The algorithms are often expensive in terms of training time and cannot scale to large, web-scale datasets.

We address both these issues by developing MIML\textsuperscript{perf}, a plug-in classifier for the MIML problem. MIML\textsuperscript{perf} directly tries to optimize complex performance measures such as macro and micro-F measure in a scalable manner. MIML\textsuperscript{perf} excels over its competitors in its ability to predict rare labels correctly. A notable feature of MIML\textsuperscript{perf} is its streaming nature, that allows it to be executed by making several passes over the data without the need for storing the entire dataset in memory. We rigorously benchmark MIML\textsuperscript{perf} to establish that 1) it offers far greater label extraction accuracies on RE than specialized methods for the problem, 2) it also outperforms state-of-the-art MIML approaches on text categorization and scene classification problems, and 3) it can offer orders of magnitude faster running times.

2 Related Work
Completely supervised approaches, such as those for RE (Zhou et al. 2005; Surdeanu and Ciaramita 2007) have limited application owning to their requirement of fully annotated data which is expensive. Of the alternate paradigms, that of distant supervision has found much appeal in both RE and MIML literature. For the RE problem, this paradigm shift due to (Mintz et al. 2009) fueled much interest with

\textsuperscript{*}Work done while the author was a student at IIT Bombay

Copyright \textcopyright{} 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
and Zhan 2006) provides two methods, viz. (Zeng et al. 2015; Zhan and Zhou 2008) that have been applied to text pose algorithms for the MIML problem (Zhou and Zhan 2012), their approach suffers from two drawbacks: (i) the heavy tail of rare classes with small class priors. (ii) the approach does not perform well for the datasets and (ii) the approach does not perform well for the multi-instance single label (SIML) problems and then converts each into Single Instance Single Label (SISL) problems. MIMLSVM, on the other hand, converts the MIML problem into smaller independent sub-problems over the large unlabeled corpus. The training data consists of entity pairs in the database to their mentions in the corpus, in other words, a multi-instance single-label learning problem. Soon, (Hoffmann et al. 2011; Surdeanu et al. 2012) generalized this to allow entity pairs to participate in multiple relations, thus completing the MIML abstraction. A variety of techniques have been applied to this problem over the years. Recently (Zeng et al. 2015) applied Piecewise Convolutional Neural Networks (PCNNs) to the RE problem. However, these models were often trained by optimizing performance measures such as conditional log-likelihood (Surdeanu et al. 2012), error rate, or bag level entropy (Zeng et al. 2015) that are not directly related to the measures actually used for evaluation such as F-measure and average precision.

The area of learning structured prediction models becomes relevant in this context since MIML requires predicting a structured array of labels. The area has seen the development of powerful large-margin methods (Taskar, Guestrin, and Koller 2003) which can incorporate hidden variables (Wang and Mor 2011; Felzenszwalb et al. 2010; Yu and Joachims 2009), which is useful since MIML admits elegant formulations as a latent variable learning problem, as well as optimize non-decomposable performance measures such as F-measure (Ranjbar et al. 2013; Rosenfeld et al. 2014; Keshet 2014).

However, the only direct application of these techniques to optimize non-decomposable performance measures in the MIML setting is in the work of Haffari et al. (2015), who optimize the micro averaged F-measure in the RE setting. Their optimization algorithm interleaves the Concave-Convex Procedure (CCCP) (Yuille and Rangarajan 2001) to populate latent variables using dual decomposition (Rush and Collins 2012). This factorizes the hard optimization problem into smaller independent sub-problems over the training instances. Despite such optimization tricks and the use of heuristic local search methods replacing the exhaustive search of (Joachims 2005; Ranjbar, Vahdat, and Mori 2012)), their approach suffers from two drawbacks: (i) the resulting algorithms are slow and do not scale to large datasets and (ii) the approach does not perform well for the heavy tail of rare classes with small class priors.

There also has been progress in developing general purpose algorithms for the MIML problem (Zhou and Zhan 2006; Zhan and Zhou 2008) that have been applied to text and image (scene) classification tasks. The work of (Zhou and Zhan 2006) provides two methods, viz., MIMLBoost and MIMLSVM for the problem. MIMLBoost solves several multi-instance single label (MISL) problems and then converts each into Single Instance Single Label (SISL) problems using multi-instance boosting. MIMLSVM, on the other hand, converts the MIML problem into several Single Instance Multi Label (SIML) problems using k-medoids clustering that uses the Hausdorff distance between all pairs of points and then solves the SIML problems using SVM. Both methods work on the degenerate versions of MIML and hence can be lossy. The main bottleneck in the MIMLSVM approach is the extremely expensive computation of Hausdorff distance matrix between all pairs of points. As we shall see in Section 5, this causes this approach to be extremely slow. As a concluding note, we point out that the the SIML approach of reducing the problem of predicting labels for bags to that of predicting labels for individual instances is very popular in RE (Zeng et al. 2015; Haffari, Nagesh, and Ramakrishnan 2015; Hoffmann et al. 2011) and MIML algorithms. We will revisit this approach while describing the MIML$^{perf}$ algorithm.

The principles of EM, alternating minimization, etc., are very popular in optimization and learning literature. We note that the classical EM algorithm cannot be directly applied to optimize multivariate performance measures as it is designed to optimize the decomposable likelihood measure over i.i.d. samples. As a result, direct adaptations of EM (Zhang and Goldman 2001) to multi-instance settings only address decomposable measures and that too, only in binary or multi-class settings where labels are exclusive. The models developed therein cannot extend directly even to decomposable performance measures in the multi-label setting since labels can co-exist in this setting, something that our approach critically exploits.

### 3 Problem Formulation

In the MIML setting, training data is presented as $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^{N}$ where $x_i = \{x_i^{(1)}, \ldots, x_i^{(n_i)}\} \in \mathcal{X}$ is the $i^{th}$ instance set (also called a bag) containing $n_i$ instances and $y_i = [y_{i,1}, y_{i,2}, \ldots, y_{i,L}] \in \mathcal{Y} = \{0, 1\}^L$ is a vector of labels associated with $x_i$. We will use $h_i \in \{1, \ldots, L, \text{nil}\}^{n_i}$ to denote the vector of hidden labels for $x_i$, $h_{i,j}$ will encode whether the $j^{th}$ instance in the bag $x_i$ expresses one of the labels $\{1, \ldots, L\}$ or none (nil). We will denote $\mathbf{X} := \{x_i\}_{i=1}^{N} \in \mathcal{X}$ and $\mathbf{Y} := \{y_i\}_{i=1}^{N} \in \mathcal{Y}$.

Let us illustrate the above with an example from the RE domain. The goal here is to align facts to sentences in a large unlabeled corpus. The training data consists of entity pairs, such as (Walt Disney, Mickey Mouse). For each entity pair, we are given a set of sentences (also called mentions) that talk about these entities, and a set of relations known to be satisfied by the pair. For instance, (Walt Disney, Mickey Mouse) satisfy the relations Co-creator-Of, Voice-Of but not the relation Animator-Of (Mickey was animated by Ub Iwerks).

The above can be cast as an MIML problem by letting $x_i$, the instance set or bag containing all mentions of the $i$-th entity pair and $y_i$ be the relations satisfied by the pair. The hidden label vector $h_i$ can then be used to denote relations expressed by individual mentions. The set of $L$ possible relations is extracted from a knowledge-base.

We wish to learn a function $f: \mathcal{X} \to \mathcal{Y}$ that predicts labels for novel data points $f(x) = [f_1(x), f_2(x), \ldots, f_L(x)]$. Our learning process will be guided by a performance measure $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$. Given a dataset $(X, Y)$ containing $N$ bags, let $Y = [y_{1,j}, y_{2,j}, \ldots, y_{N,j}]$ encode which bags support label $j$ and $P(X) = [f_1(x_1), f_2(x_2), \ldots, f_L(x_N)]$ be the vector of predictions for label $j$.

The simplest of performance measures is the class of univariate or decomposable measures that compute performance over a set of data points by simply averaging
the performance on individual data points \( \Delta(f(X), Y) = \frac{1}{N} \sum_{i=1}^{N} \Delta(f(x_i), y_i) \). Examples include the Hamming distance (Bi and Kwok 2013; Chen and Lin 2012), precision (Hsu et al. 2009; Weston, Bengio, and Usunier 2011), and recall (Steck 2010). Although convenient to work with and analyzable, these performance measures are known to be ill suited in the presence of label imbalance or a heavy tailed label distribution (Koyejo et al. 2014; Narasimhan, Vaish, and Agarwal 2014) since they tend to neglect performance on rare labels. Incidentally, the heavy tail phenomenon is well documented in relation extraction settings (see Figure 1a).

Instead, the performance measures of choice in these situations are the multivariate performance measures since they force the predictor to do well on rare labels as well. These measures are typically non-decomposable as their evaluation does not decompose over individual points. In this paper we consider the family of F-measures – these have been popular for the RE problem (Haffari, Nagesh, and Ramakrishnan 2015; Surdeanu et al. 2012), as well as for label imbalanced learning in general (Koyejo et al. 2014; Ye et al. 2012).

The label-wise Precision and Recall of a predictor \( f \) measure the performance on each individual label

\[
\text{PREC}^j(f; X, Y) := \frac{\sum_{i=1}^{n} y_{i,j} f^j(x_i)}{\sum_{i=1}^{n} y_{i,j}}
\]

\[
\text{REC}^j(f; X, Y) := \frac{\sum_{i=1}^{n} y_{i,j} f^j(x_i)}{\sum_{i=1}^{n} y_{i,j}}
\]

whereas the global Precision and Recall calculate the overall performance of \( f \) across labels

\[
\text{PREC}(f; X, Y) := \frac{\sum_{i=1}^{n} \sum_{j=1}^{L} y_{i,j} f^j(x_i)}{\sum_{i=1}^{n} \sum_{j=1}^{L} y_{i,j}}
\]

\[
\text{REC}(f; X, Y) := \frac{\sum_{i=1}^{n} \sum_{j=1}^{L} y_{i,j} f^j(x_i)}{\sum_{i=1}^{n} \sum_{j=1}^{L} y_{i,j}}
\]

For any label \( j \), we define label-wise F-measure as

\[
F^j_\beta(f; X, Y) := \left( \frac{\beta \cdot \text{PREC}^j(f; X, Y)}{1 - \beta} + \frac{1 - \beta}{\text{REC}^j(f; X, Y)} \right)^{-1}
\]

as well as the Macro and Micro F-measure as

\[
F^\text{macro}_\beta(f; X, Y) := \frac{1}{L} \sum_{j=1}^{L} F^j_\beta(f; X, Y),
\]

\[
F^\text{micro}_\beta(f; X, Y) := \left( \frac{\beta \cdot \text{PREC}(f; X, Y)}{1 - \beta} + \frac{1 - \beta}{\text{REC}(f; X, Y)} \right)^{-1}
\]

where \( \beta = 0.5 \) gives the standard F1 measure. These performance measures, especially the macro F-measure, penalize predictors that perform poorly on rare labels. In the next section, we develop scalable techniques for optimizing multivariate performance measures such as F-measure variants in the MIML and RE settings.

4 Proposed Approach

Despite having attractive properties as discussed above, multivariate performance measures such as F-measure present challenges to learning algorithms. Due to their non-decomposability, classical algorithmic tools such as online and stochastic optimization, as well as analytical tools such as uniform convergence bounds are not readily applicable.

The recent years have seen a growing interest in the problem of optimizing non-decomposable performance measures, especially F-measure variants (Dembczynski et al. 2013; Narasimhan, Kar, and Jain 2015) However, past work is mostly restricted to binary classification. An exception is the recent work of Haffari et al. (2015) that utilizes the structural SVM approach (Joachims 2005) to optimize the F-micro measure in the RE setting. However, the approach is woefully non-scalable, struggling to cope with even a few thousand data points. This severely restricts its applicability to real life and production-grade problems.

This section will develop MIML\textsuperscript{perf}, a scalable tool for optimizing multivariate performance measures in the MIML setting. MIML\textsuperscript{perf} is based on a plug-in approach to classification. Plug-in classifiers have been studied in binary classification settings (Kotłowski and Dembczynski 2015; Narasimhan, Vaish, and Agarwal 2014; Koyejo et al. 2014; Ye et al. 2012) with great success. However, to the best of our knowledge, plug-in approaches have not been studied in MIML/RE settings before.

Plug-In Classifiers

Consider a simple binary classification problem where the task is to assign every data point \( x \in \mathcal{X} \), a binary label \( y \in \{\pm 1\} \). Plug-in classifiers achieve this by first learning to predict Class Probability Estimate (CPE) scores. A function \( g: \mathcal{X} \rightarrow \mathbb{R} \) is learnt such that \( g(x) \approx P[y = 1] \). Various tools such as logistic regression may be used to learn this CPE model \( g \). The final classifier is of the form

![Diagram](Image)
Algorithm 1 MIML$^{perf}$: Testing Routine

\textbf{Input}: Data $\{\{x_i, y_i\}\}_{i=1}^N$, expression rate $\kappa$, perf. measure $\Delta$
1: For all $(i, j)$ such that $y_{i,j} = 1$, set random $\kappa \cdot n_i$ entries of
2: the vector $z^{(i,j)}$ to 1 // Initialize hidden labels randomly
3: while not converged do
4: \hspace{1em} Step 1: Fix hidden variables, update plug-in classifiers
5: \hspace{2em} for every label $j \in [L]$ do
6: \hspace{3em} $D' \leftarrow \{(x^{(k)}, z^{(i,j)})\}_{i=1}^{n_i}$ // Prepare datasets
7: \hspace{3em} $g' \leftarrow \text{CPE-train}(D')$ // Train CPE models
8: \hspace{2em} end for
9: \hspace{1em} end while
10: $\eta \leftarrow \text{Tune-thresholds}((X, Y), p; \Delta)$ // Optimize $\Delta$

\textbf{Step 2}: Fix plug-in classifiers, update hidden variables
8: for $(i, j) \in [N] \times [L]$ such that $y_{i,j} = 1$ do
9: $z^{(i,j)} \leftarrow 0^{n_i}$ // Reset hidden labels
10: $c^{(i,j)} \leftarrow \sum_{k=1}^{n_i} 1\{g'(x^{(k)}) \geq \eta_j\}$ // Discover hidden labels
11: $S^{(i,j)} \leftarrow \text{Sample } c^{(i,j)} \text{ entries of } z^{(i,j)} \text{ according to } g'$
12: $z^{(i,j)} \leftarrow 1$ for all $k \in S^{(i,j)}$ // Reestimate hidden labels
13: end for
14: end for

Algorithm 2 MIML$^{perf}$: Training Routine

\textbf{Input}: Test point $x = \{x^{(1)}, \ldots, x^{(n_t)}\}$ with $n_t$ instances, CPE models $g_1, \ldots, g_L$, thresholds $\eta = [\eta_1, \ldots, \eta_L]$
1: for $k = 1, 2, \ldots, n_t$ do
2: $h_k \leftarrow \{j : g^j(x^{(k)}) \geq \eta_j\}$ // Discover hidden labels
3: end for
4: for $j = 1, 2, \ldots, L$ do
5: $y_j \leftarrow \bigvee_{k=1}^{n_t} 1\{j \in h_k\}$ // Aggregation step
6: end for
7: return $\hat{y} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_L]$

$\text{sign}(g(x) - \eta)$ where $\eta$ is a threshold that is tuned to maximize the performance measure being considered, e.g. classification accuracy, F-measure, G-mean etc.

Plug-in approaches offer various benefits. They can be used to optimize complex multivariate performance measures. In fact the same CPE model can be reused to target several performance measures by simply changing the threshold tuning step. Plug-in approaches have been rigorously analyzed and are known to be statistically consistent (Narasimhan, Vaish, and Agarwal 2014).

However, plug-in methods require total supervision, akin to requiring every instance in every bag to be labeled, whereas MIML operates under a significantly impoverished distant supervision setting. Thus, a direct application of existing plug-in approaches to MIML problems is not possible.

**MIML$^{perf}$: An MIML Algorithm for Optimizing Multivariate Performance Measures**

MIML$^{perf}$ reduces the problem of predicting labels for a bag to the problem of predicting labels expressed by individual instances in that bag. Recall that this is captured by the hidden variables (see Section 3). E.g. the sentence “Mickey Mouse was voiced by Walt Disney himself” expresses the label Voice Of but not the label Co-creator Of. Afterward, a label is predicted as relevant for the bag if at least one instance in the bag expresses that label. This aggregation or SIML approach has been standard in MIML and RE settings (Haffari, Nagesh, and Ramakrishnan 2015; Hoffmann et al. 2011; Zhan and Zhou 2008).

Existing approaches suffer since they use expensive techniques such as integer linear programming Haffari et al. (2015) to set these hidden variables. MIML$^{perf}$ instead uses plug-in classifiers to set the hidden variables which results in training routines that are orders of magnitude faster. However, training these classifiers is challenging since we do not have any instance level supervision — the training data does not tell us which instances express which labels.

**Overview of MIML$^{perf}$ Training:** MIML$^{perf}$ overcomes the challenge of absence of instance level supervision in a scalable manner by combining two powerful approaches — plug-in classifiers and alternating optimization. At every time step, MIML$^{perf}$ estimates the hidden variables, i.e. for each instance in a training bag, it estimates which labels are likely to be expressed by that instance. These estimates are valuable since they offer a form of instance level supervision to the method. This makes it possible for MIML$^{perf}$ to train plug-in classifiers by learning CPE models and tuning appropriate thresholds; a separate classifier is trained for each of the $L$ labels. MIML$^{perf}$ uses these classifiers to get improved estimates of the hidden variables. The process is repeated for a few iterations. Details of the training and testing routines for MIML$^{perf}$ are given in Algorithms 1 and 2.

**Hidden Variables** These are used to record the algorithm’s beliefs about which labels are expressed by individual instances. For a set of $N$ data points $\{(x_i, y_i)\}_{i=1}^N$, the binary variable $z^{(i,j)} \in \{0, 1\}$ indicates whether the $k$th instance in the $i$th bag expresses the $j$th label or not. If the $i$th data point has $n_i$ instances then the vector $z^{(i,j)} \in \{0, 1\}^{n_i}$ encodes which of the instances in the bag express the label $j$. A crucial observation that makes our approach scalable is that if a bag does not have a label $j$, then none of its instances can express it. In other words, if $y_{i,j} = 0$ then $z^{(i,j)} = 0$. This observation is valuable since we need only worry about $z^{(i,j)}$ when $y_{i,j} = 1$. Since most bags have very few labels (see Figure 1a), this introduces sparsity into our approach. We now describe the alternation steps.

**Step 1: Learning Plug-In Classifiers** Given an assignment to all the hidden variables $z^{(i,j)}$, MIML$^{perf}$ learns a plug-in classifier for each label. The job of these classifiers corresponding to a particular label is to predict whether a given instance expresses that label or not. This is done by simply formulating a binary classification problem for each label with the hidden variables acting as the “classes”. A standard procedure like logistic regression (CPE-train in Algorithm 1) is used to obtain CPE models, one for each label. Afterward, thresholds are tuned for each label such that the performance measure $\Delta$ being targeted, such as F-macro or F-micro measure, is maximized. Doing so in an efficient manner is challenging as in large scale settings, the number of CPE scores being handled can run into millions or more.\footnote{We refer the reader to https://www.cse.iitb.ac.in/~ganesh/papers/supp-aaai17/ for details.}

**Step 2: Re-estimating Hidden Variables** The CPE mod-
els learnt in the previous step can be used to obtain (noisy) CPE scores for all instances with respect to all labels. The noise is due to incorrect assignments to the hidden variables in the previous step. Nevertheless, this gives us an estimate of which instances are more likely to express a certain label. MIML$^\text{perf}$ uses these CPE scores to reassign the hidden variables. In order to avoid trusting these scores completely, MIML$^\text{perf}$ chooses a random set of instances with high CPE scores with respect to a label and assigns them that label. This has the effect of reinforcing good CPE models as well as smoothing out errors. For the initialization step, if a bag has a label $j$, the method assigns a random $\kappa$ fraction of instances in that bag to label $j$. $\kappa$ is an expression rate parameter that is only used for initialization.

**Overview of MIML$^\text{perf}$** Training: The testing procedure of MIML$^\text{perf}$ (see Algorithm 2) involves simply applying the plug-in classifiers and performing the aggregation step to obtain the predictions for all labels for a bag. This procedure is enormously cheaper than the integer linear programming approach followed by Haffari et al. (2015).

**Theoretical Analysis**

Giving strong theoretical guarantees for algorithms for non-convex learning problems such as multi-instance learning is a challenging albeit interesting problem. We are nevertheless able to establish generalization guarantees for our method which proves that our method does not overfit. Let $\hat{\mathbf{f}}$ be the classifier generated by MIML$^\text{perf}$. Let the training and test sets $(\mathbf{X}, \mathbf{Y})$ and $(\mathbf{X}^t, \mathbf{Y}^t)$ be chosen randomly from some fixed but unknown distribution. Let $\pi$ denote the minimum frequency of any label.

**Theorem 1.** Let the instances be represented as $d$ dimensional features $\mathbf{x}^{(k)} \in \mathbb{R}^d$. Then for any $N$ such that

$$\sqrt{\frac{1}{N} \left( \log \frac{1}{\delta} + d \log \frac{N}{\delta} \right)} < c \cdot \pi$$

for a small constant $c$, we have with probability at least $1 - \delta$, for some constant $C$,

$$\mathbb{E} \left[ F^{\text{macro}}_\beta (\hat{\mathbf{f}}; \mathbf{X}^t, \mathbf{Y}^t) - F^{\text{macro}}_\beta (\hat{\mathbf{f}}; \mathbf{X}, \mathbf{Y}) \right] \leq \sqrt{\frac{c}{N} (d + \log \frac{1}{\delta})}$$

The result is stated for linear models for sake of simplicity and can be extended to hypothesis spaces with finite capacity as well. A similar result holds for F-micro measure as well.

**5 Experiments**

We present detailed comparisons of our approach with the state of the art on three benchmark MIML/RE datasets.

**1. Riedel Distant Supervision Dataset:** For the distantly supervised relation extraction problem, we use the benchmark dataset created by (Riedel, Yao, and McCallum 2010). The dataset was created by aligning relations from Freebase with the sentences in the New York Times corpus (Sandhaus 2008). The labels for the data points come from the Freebase database; however, Freebase is incomplete (Ritter et al. 2013). A data point is labeled nil when either no relation exists or the relation is absent in Freebase. Following (Haffari, Nagesh, and Ramakrishnan 2015), we train and evaluate all algorithms on a subset of this dataset, termed as the positive dataset, which consists of only non-nil relation labeled data points.

**2. MIML Scene Classification Dataset (Scene):** The Scene data set contains 2000 scene images collected from the COREL image collection and the Internet, with five different possible class labels, viz., desert, mountains, sea, sunset and trees. Each image is represented as a bag of nine instances with each instance represented as a 15-dimensional feature vector that corresponds to an image patch. Following MIMLSVM, We divided the data into two parts consisting of 1600 data points for training and the remaining 400 points for testing.

**3. MIML Text Classification Dataset (Reuters):** The text data is derived from the widely studied Reuters-21578 collection using seven most frequent classes. It consists of 2000 documents with 15% associated with more than one class labels. In this dataset each instance is represented as a 243-dimensional feature vector and corresponds to a text segment. Again, we follow MIMLSVM in the way we partition this dataset into training and testing splits.

**Comparisons against existing approaches**

We compare MIML$^\text{perf}$ against state-of-the-art methods for each of the three datasets. We report $F^{\text{macro}}_\beta$, $F^{\text{micro}}_\beta$ and average precision as our evaluation measures. In addition, we also report the training time and the average time required to predict labels for each test bag (instance set). We used a default value $\kappa = 1$ as the prevalence parameter for training MIML$^\text{perf}$. For the sake of fairness, training splits and feature sets were kept common across all comparative experiments.

<table>
<thead>
<tr>
<th>Table 1: Dataset Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>#bags</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Riedel</td>
</tr>
<tr>
<td>Scene</td>
</tr>
<tr>
<td>Reuters</td>
</tr>
</tbody>
</table>

Table 2 compares MIML$^\text{perf}$ against state-of-the-art approaches MIMLSVM and M3MIML (Zhan and Zhou 2008) for the Scene and Reuters datasets. We observe that MIML$^\text{perf}$ consistently outperforms both methods on all three evaluation measures. Additionally, as noted in Table 5, MIML$^\text{perf}$ is significantly and consistently faster in training than MIMLSVM and even more so at testing time.

Table 3 compares MIML$^\text{perf}$ on the distant supervision RE setting with the MM-F$_\beta$ approach (Haffari, Nagesh, and Ramakrishnan 2015). We note that MIML$^\text{perf}$ offers significant improvements over MM-F$_\beta$ on F-macro along with a 145-fold speedup in training. The gains on F-micro are marginal. However, in skewed class distribution settings, F-macro is a more relevant measure of performance than F-micro.

In addition to offering gains on training and test speeds, our approach does not require the entire dataset to be loaded in memory. As noted earlier, MIML$^\text{perf}$ is executed by making several passes over the data and is inherently streaming in nature. None of the other approaches we have compared against, enjoy these properties.
Table 2: Performance of $\text{MIML}^{perf}$ compared against MIMLSVM and M3MIML on the Scene and Reuters Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Kernel</th>
<th>F-Macro</th>
<th>F-Micro</th>
<th>Avg. Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MIMLSVM</td>
<td>MIMML</td>
<td>$\text{MIML}^{perf}$</td>
<td>MIMLSVM</td>
</tr>
<tr>
<td>Scene</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reuters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: MIML$^{perf}$ comparison with the (MM-F$^\beta$) model by Haffari et. al. on the Riedel dataset

<table>
<thead>
<tr>
<th></th>
<th>F-macro</th>
<th>Precision</th>
<th>Recall</th>
<th>F-micro</th>
<th>Train time</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM-F$^\beta$</td>
<td>0.1366</td>
<td>0.6599</td>
<td>0.6521</td>
<td>0.6559</td>
<td>4h 20m</td>
</tr>
<tr>
<td>$\text{MIML}^{perf}$</td>
<td>0.2283</td>
<td>0.7928</td>
<td>0.5781</td>
<td>0.6686</td>
<td>1m 49s</td>
</tr>
</tbody>
</table>

Table 4: Effect of the training objective on test performance on the Riedel dataset

<table>
<thead>
<tr>
<th></th>
<th>F-macro</th>
<th>F-micro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimizing $F^{macro}_\beta$</td>
<td>22.8354</td>
<td>61.0615</td>
</tr>
<tr>
<td>Optimizing $F^{macro}_\beta$</td>
<td>18.984</td>
<td>66.8648</td>
</tr>
</tbody>
</table>

Table 5: Comparisons of training time and testing time per instance

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Kernel</th>
<th>Training time</th>
<th>Speedup</th>
<th>Testing time per instance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MIMLSVM</td>
<td>$\text{MIML}^{perf}$</td>
<td>MIMLSVM</td>
</tr>
<tr>
<td>Scene</td>
<td>Linear</td>
<td>7m 45s</td>
<td>1.66s</td>
<td>280×</td>
</tr>
<tr>
<td></td>
<td>RBF</td>
<td>9m 24s</td>
<td>6m 40s</td>
<td>1.41×</td>
</tr>
<tr>
<td>Reuters</td>
<td>Linear</td>
<td>3m 55s</td>
<td>1.62s</td>
<td>145×</td>
</tr>
<tr>
<td></td>
<td>RBF</td>
<td>5m 04s</td>
<td>1m 02s</td>
<td>5×</td>
</tr>
</tbody>
</table>

Table 6: Relation-wise comparisons of F-score with MM-F$^\beta$, $\text{MIML}^{perf}$, $\kappa = 1$ and $\text{MIML}^{perf}$, tuned $\kappa$ on Riedel

<table>
<thead>
<tr>
<th>Relation</th>
<th>% of occurrence</th>
<th>MM-F$^\beta$</th>
<th>$\text{MIML}^{perf}$, $\kappa = 1$</th>
<th>$\text{MIML}^{perf}$ tuned $\kappa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>/people/person/place_lived</td>
<td>12.36</td>
<td>0.15966</td>
<td>0.46856</td>
<td>0.46897</td>
</tr>
<tr>
<td>/people/deceased_person/place_of_death</td>
<td>4.04</td>
<td>0.19231</td>
<td>0.31111</td>
<td>0.35294</td>
</tr>
<tr>
<td>/location/administrative_division/country</td>
<td>1.26</td>
<td>0</td>
<td>0.37624</td>
<td>0.37624</td>
</tr>
<tr>
<td>/business/company/founders</td>
<td>1.09</td>
<td>0.48148</td>
<td>0.68571</td>
<td>0.75676</td>
</tr>
<tr>
<td>/location/country/capital</td>
<td>0.77</td>
<td>0</td>
<td>0.10866</td>
<td>0.11494</td>
</tr>
<tr>
<td>/film/film/featured_film_locations</td>
<td>0.15</td>
<td>0</td>
<td>0.01262</td>
<td>0.01325</td>
</tr>
<tr>
<td>Overall F-Macro</td>
<td></td>
<td>0.13660</td>
<td>0.22835</td>
<td>0.23282</td>
</tr>
</tbody>
</table>

Effects of changes in parameters and objective In Table 4 we compare the difference in performances of $\text{MIML}^{perf}$ when the macro-F measure is replaced with the micro-F measure as the objective. We compare the results of training to optimize the $F^{\text{macro}}_\beta$ measure against optimizing $F^{\text{micro}}_\beta$. In each setting, we report both $F^{\text{micro}}_\beta$ and $F^{\text{macro}}_\beta$. We note the strong correspondence between the choice of performance in the training objective and performance being evaluated upon the test dataset.

We expect optimizing the F-macro performance measure to enhance performance on the rarer classes (that have low priors). We confirm this in Table 6 which compares the performance of $\text{MIML}^{perf}$ with MM-F$^\beta$ for some of the rarer classes. Whereas MM-F$^\beta$ reports extremely low F-scores (in some cases 0) on these rare classes, $\text{MIML}^{perf}$ performs much better, even with $\kappa = 1$. The value of $\kappa$ yielding the best performance for a class changes across classes and therefore, also report results with $\kappa$ tuned for each class on the training dataset. This gives us even better gains.

We also test for statistical differences between the accuracies of $\text{MIML}^{perf}$ and MM-F$^\beta$ on all the 29 classes using the Wilcoxon$^3$ signed-rank test. We notice that the sum of the signed ranks (222 with $\kappa = 1$ and 237 for $\kappa$ tuned for each class) is very clearly in favour of $\text{MIML}^{perf}$ over MM-F$^\beta$.

6 Conclusions

We proposed $\text{MIML}^{perf}$, a scalable approach for optimizing several multivariate performance measures in the MIML setting and applied it to relation extraction and other MIML tasks. Our approach makes novel use of the plug-in technique, offers significantly reduced running times, and caters well even to rare classes, without compromising on the popular ones. Furthermore, our approach appears very amenable to distributed and parallel computing and a possible future work is validating this hypothesis.

Acknowledgements

PK thanks the Deep Singh and Daljeet Kaur Faculty Fellowship and the Research-I Foundation, IIT Kanpur for support.

$^3$The signed ranks for the 29 classes of the Riedel dataset are available as a spreadsheet at https://www.cse.iitb.ac.in/~ganesh/papers/supp-aaai17/.
References


