Active Learning for Sparse Bayesian Multilabel Classification

Deepak Vasisht
MIT
deepakv@mit.edu

Andreas Damianou
University of Sheffield, UK
andreas.damianou@sheffield.ac.uk

Manik Varma
Microsoft Research
manik@microsoft.com

Ashish Kapoor
Microsoft Research
akapoor@microsoft.com

ABSTRACT

We study the problem of active learning for multilabel classification. We focus on the real-world scenario where the average number of positive (relevant) labels per data point is small leading to positive label sparsity. Carrying out mutual information based near-optimal active learning in this setting is a challenging task since the computational complexity involved is exponential in the total number of labels. We propose a novel inference algorithm for the sparse Bayesian multilabel model of [17]. The benefit of this alternate inference scheme is that it enables a natural approximation of the mutual information objective. We prove that the approximation leads to an identical solution to the exact optimization problem but at a fraction of the optimization cost. This allows us to carry out efficient, non-myopic, and near-optimal active learning for sparse multilabel classification. Extensive experiments reveal the effectiveness of the method.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Performance, Machine Learning, Optimization

Keywords

Multi-label Learning; Active Learning; Mutual Information

1. INTRODUCTION

The goal in multilabel classification is to learn a classifier which can automatically tag a data point with the most relevant *set* of labels. This is in contrast to multi-class classification where only a single label needs to be predicted per data point. Our objective, in this paper, is to develop an efficient, non-myopic and near-optimal active learning algorithm for multilabel classification employing a mutual information based data point selection criterion.

KDD'14, August 24–27, 2014, New York, NY, USA.

ACM 978-1-4503-2956-9/14/08 ...$15.00.
http://dx.doi.org/10.1145/2623330.2623759 .

There has been much recent interest in sparse multilabel learning [2, 15, 17]. In this scenario, each data point exhibits positive label sparsity in that only a small fraction of any point's labels are ever marked as positive or relevant. Note that this is a typical setting in most real world applications – the average number of positive labels per data point ranges from less than 5% on most UCI data sets to less than 0.001% on Wikipedia and other web data sets [2]. Sparse multilabel learning therefore presents an important, real-world and novel challenge to active learning techniques.

Traditional active learning techniques for multilabel learning do not take positive label sparsity into account [11, 13, 20,21,31]. Furthermore, these techniques are myopic and actively select only a single data point at a time for annotation. As a result, there is no provable guarantee that the final set of annotated points is optimal or even near-optimal. While mutual information based data sampling can provide near-optimal guarantees in certain regression settings [18,19], it is non-trivial to extend those to the active multilabel setting.

In this paper, we first develop an alternate inference technique for the sparse Bayesian multi-label graphical model of [17] which allows us to carry out efficient mutual information based active learning. We develop an approximation to the mutual information which is tightly coupled to our inference algorithm and which can be optimized much more efficiently than the exact mutual information. This approximation allows us to develop an active learning strategy that has the following desirable properties:

- The set of annotated points selected by our algorithm is provably near optimal.
- The selection of annotated points can be done in time independent of the size of the label space if all labels of the point are to be annotated, as opposed to the existing methods which scale linearly (at best) with the number of labels.
- The train and test times are considerably reduced due to the reduction of the label vector to a low dimensional space using a compressed sensing matrix.
- The method can be easily adapted to be integrated with different compressed sensing approaches in literature [3, 5, 8, 10, 12, 15–17, 28, 29, 32, 33].
- The method takes into account label sparsity which has been proven to be a desirable characteristic in past work on multilabel classification [2, 15, 17].

- The method allows us to generalize to other active learning scenarios like active diagnosis (selecting which labels to annotate for a point) and generalized active learning (selecting both the points and the labels to annotate).

Our main technical contribution is Theorem 1 where we prove that, at any given data point selection step, our approximation leads to an *identical* solution as would have been obtained using exact mutual information in the limiting case, but at a fraction of the optimization cost. Furthermore, by performing inference in a sparse Gaussian Process regression model, we can leverage the mutual information guarantees to provably show that the entire subset of data points selected for annotation by our proposed algorithm is near-optimal. This enables us to carry out active sparse multilabel learning in an efficient, non-myopic and theoretically principled fashion. Finally, as a natural extension of our method, the matrix inversions necessary for traditional active learning can be carried out in time that is independent of the size of the label vector for a point.

Extensive experiments reveal that our proposed active learning strategy consistently outperforms the state-of-the-art approaches based on variance sampling, SVM based active learning proposed in [20] and random sampling baselines. Furthermore, we demonstrate that our inference procedure leads to a better estimate of the variance needed for Bayesian active learning as compared to existing techniques. We also show that our method has significant gains (20 times gain over the `delicious` dataset with 983 labels) in time consumption over the state-of-the-art [20] as the number of labels increases. Finally, we demonstrate that our proposed approach can also be used to sample both labels as well as data points, thereby allowing learning at an even lower annotation cost as compared to existing techniques which have so far focussed on all the labels being annotated for a selected data point.

Our main contributions are as follows: (a) a novel inference procedure for the Bayesian sparse multilabel graphical model of [17]; (b) an approximation to the mutual information which follows naturally from our proposed inference procedure; (c) a proof that our proposed approximation selects the same data point as would have been chosen by optimizing mutual information (in the limiting case); and (d) an extensive evaluation of the proposed scheme on a diverse range of real world datasets to prove the gains obtained by our algorithm. These contributions allow us to carry out efficient mutual information based active sparse multilabel learning for the first time (to the best of our knowledge).

2. RELATED WORK

Past work in active learning has primarily focused on single label classification tasks [24]. However, multilabel classification has received wide interest in recent times [2, 3, 5, 8, 10, 12, 15–17, 28–30, 32, 33]. Active learning is more advantageous in this setting as label acquisition costs are higher for multilabel scenario. However, research in active learning for multilabel classification is still in its preliminary stage. Current multilabel active learning strategies train a binary SVM classifier for each label and combine the SVM margins using different heuristics to select the training set from a pool of available data for annotation [11, 21, 27]. [27] takes the average of the SVM uncertainties obtained by Platt scaling [23] and uses it as a selection criterion for greedily selecting the set of points to be annotated. [21] uses Mean Max Loss(MML) and Max Loss(ML) as two separate selection criteria for annotation of points. [31] uses logistic regression to predict the number of positive labels for each point and uses another SVM based heuristic to minimize expected loss for point selection. The state-of-the-art method [20] uses a combination of deviation from mean label cardinality and an SVM-based loss function for active learning. To the best of our knowledge, all these works are myopic and do not provide any theoretical guarantees on the optimality of the selected set.

On the other hand, mutual information has proved to be a very useful criterion for active learning with Gaussian Processes [18,19,26], both empirically and theoretically. In these schemes, each point is selected from a pool of available unlabeled data, so as to maximize the information gain over the remaining unlabeled data. The approximate submodularity of mutual information ensures that a set of points selected greedily based on this criterion will be near-optimal [19]. However, extending mutual information to multilabel classification is non-trivial as a straightforward computation of mutual information over the label space is exponential in the number of labels.

The Bayesian compressed sensing model proposed in [17] is closest to our work, in the sense that it models the multilabel classification task as a Gaussian Process. However, the approximations made in [17], for the inference procedure to be tractable, do not preserve the covariance matrix across labels and hence, render mutual information based active learning infeasible. The inference procedure proposed in section 4, however, preserves this covariance matrix and allows for efficient mutual information based active learning.

3. BACKGROUND

Our work builds upon a recently proposed Bayesian architecture for multilabel classification [17]. We briefly summarize this framework here. Given a set of N data points $\mathbf{X} = \{\mathbf{x}_i\}$ and the corresponding sparse l-dimensional binary label vectors $\mathbf{Y} = \{\mathbf{y}_i\}$, the key idea was to first reduce the classification task to a lower multidimensional regression task via a random projection matrix $\mathbf{\Phi}$, and then during test time, infer the labels for an unseen point via approximate Bayesian inference.

Specifically, a factor graph corresponding to the model consists of three potential functions: (1) the first potential term $f_{\mathbf{x}_i}(\mathbf{W}, \mathbf{z}_i) = \exp[-\frac{||\mathbf{W}^T\mathbf{x}_i - \mathbf{z}_i||^2}{2\sigma^2}]$ defines a regression function from each input point $\mathbf{x}_i$ to the k-dimensional latent variable $\mathbf{z}_i$ via $\mathbf{W}$. (2) The second potential term $g_{\mathbf{\Phi}}(\mathbf{y}_i, \mathbf{z}_i) = \exp[-\frac{||\mathbf{\Phi}\mathbf{y}_i - \mathbf{z}_i||^2}{2\chi^2}]$ corresponds to the transformation of l-dimensional labels $\mathbf{y}_i$ to the real-valued compressed space $\mathbf{z}_i$ via a $k \times l$, where $k \ll l$, random projection matrix $\mathbf{\Phi}$. (3) Finally, the third potential $h_{\boldsymbol{\alpha}_i}(\mathbf{y}_i) = \prod_{j=1}^{l} N(y_i^j; 0, \frac{1}{\alpha_i^j})$ enforces label sparsity by introducing a zero mean-Gaussian prior on each of the labels with precision α_i^j, which in turn are Gamma distributed. Formally, the joint probability distribution takes the following form:

$$p(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, [\boldsymbol{\alpha}_i]_{i=1}^N | \mathbf{X}, \mathbf{\Phi}) = \frac{1}{Z} p(\mathbf{W}) \prod_{i=1}^{N} f_{\mathbf{x}_i}(\mathbf{w}, \mathbf{z}_i) g_{\mathbf{\Phi}}(\mathbf{y}_i, \mathbf{z}_i) h_{\boldsymbol{\alpha}_i}(\mathbf{y}_i) p(\boldsymbol{\alpha}_i). \quad (1)$$

Here, Z is the partition function (normalization term), $p(\mathbf{W}) = \prod_{i=1}^{K} N(\mathbf{w}_i, 0, I)$ is the spherical Gaussian prior on the linear regression functions, and $p(\boldsymbol{\alpha}_i) = \prod_{j=1}^{L} \Gamma(\alpha_i^j; a_0, b_0)$ is the product of Gamma priors on each individual label. Intuitively, the function $f_{\mathbf{x}_i}$ aligns the latent variable $\mathbf{z}_i$ with the output of the linear regression functions, and the function $g_{\boldsymbol{\Phi}}$ favors compatibility with the compressed label vector corresponding to $\mathbf{y}_i$. Finally, $h_{\boldsymbol{\alpha}_i}$ enforces sparsity over the labels.

Using $\mathbf{Y}_{\mathcal{L}}$ to denote labeled instances, inferring the exact posterior $P(\mathbf{Y}_{\mathcal{U}} \mid \mathbf{Y}_{\mathcal{L}}, \mathbf{X}, \boldsymbol{\Phi})$ over labels $\mathbf{Y}_{\mathcal{U}}$ for test points $\mathbf{X}_{\mathcal{U}}$ is prohibitive due to the product of Gaussian and non-Gaussian terms in the joint distribution. The approach in [17] resorts to variational inference by approximating the posterior over $\mathbf{Y}_{\mathcal{U}}$ as a Gaussian distribution. While this showed good performance in terms of classification accuracy, there is a significant disadvantage of using the same inference procedure for active learning tasks. The derived variational inference equations lead to loss in information about the variances between updates across the different layers of the graphical model. Specifically, the procedure approximates the posterior over $\mathbf{y}_i$ as a Gaussian, but the variance of the random variable $\mathbf{y}_i$ is completely independent of the variance of the random variable $\mathbf{z}_i$. Similarly, the variance of $\mathbf{z}_i$ is in turn independent of the variance of the random variable $\mathbf{W}$. The variance related to the Gaussian random variables is central to the use of mutual information in selective sampling tasks; hence, with the given inference procedure, mutual information cannot be optimally used as an active learning strategy.

4. OUR APPROACH

There are two key ingredients to our framework: first is an alternate approximate inference scheme that preserves variances due to the regression part of the graphical model. Given this inference scheme, the second part focuses on deriving an efficient near-optimal selective sampling strategy.

4.1 An Alternate Inference Procedure

Our goal here is to derive an inference scheme that would preserve the variances of the latent random variables and propagate correct uncertainties to $\mathbf{Y}$, thereby enabling an effective active learning strategy. The key observation here is the fact that such variances can in fact be preserved if instead of directly applying variational approximation, we first analytically integrate out the latent variables $\mathbf{Z}$ and $\mathbf{W}$ from the joint distribution (Eq. 1). Such analytic integration is feasible due to the form of potential functions $f_{\mathbf{x}_i}, g_{\boldsymbol{\Phi}}$ and the Gaussian Process prior $p(\mathbf{W})$. This results in a joint distribution over $\mathbf{Y}$ and $\boldsymbol{\alpha}$ that is a product of a Gaussian and a Gamma distribution. Formally,

$$p(\mathbf{Y}, [\boldsymbol{\alpha}_i]_{i=1}^N | \mathbf{X}, \boldsymbol{\Phi}) = \int_{\mathbf{Z},\mathbf{W}} p(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, [\boldsymbol{\alpha}_i]_{i=1}^N | \mathbf{X}, \boldsymbol{\Phi}) = \frac{1}{Z} e^{\frac{\mathbf{Y}^T \boldsymbol{\Sigma}_{\mathbf{Y}}^{-1} \mathbf{Y}}{2}} \prod_{i=1}^{N} h_{\boldsymbol{\alpha}_i}(\mathbf{y}_i) p(\boldsymbol{\alpha}_i).$$

Here the label vector $\mathbf{Y}$ is an Nl-dimensional vector obtained by appending all l-dimensional $\mathbf{y}_i's$ one after the other. Similarly, $\mathbf{Z}$ is an Nk-dimensional vector resulting from stacking all k-dimensional $\mathbf{z}_i's$. Further, the term $\boldsymbol{\Sigma}_{\mathbf{Y}}$ takes the following form:

$$\boldsymbol{\Sigma}_{\mathbf{Y}}{}^{-1} = \tilde{\boldsymbol{\Phi}}^T (\boldsymbol{\Sigma}_{\mathbf{Z}} + \chi^2 \mathbf{I})^{-1} \tilde{\boldsymbol{\Phi}} \quad (2)$$

$$\text{where } \boldsymbol{\Sigma}_{\mathbf{Z}} = \begin{pmatrix} K_{11}\mathbf{I}_k & K_{12}\mathbf{I}_k & . & K_{1N}\mathbf{I}_k \\ K_{21}\mathbf{I}_k & K_{22}\mathbf{I}_k & . & K_{2N}\mathbf{I}_k \\ . & . & . & . \\ K_{N1}\mathbf{I}_k & K_{N2}\mathbf{I}_k & . & K_{NN}\mathbf{I}_k \end{pmatrix} \quad (3)$$

Here, the $Nk \times Nk$-dimensional matrix $\boldsymbol{\Sigma}_{\mathbf{Z}}$ is a special block matrix, where K_{ij} is the i^{th} row and j^{th} column entry of $\mathbf{K} = (\mathbf{X}^T\mathbf{X} + \sigma^2\mathbf{I})^1$ and $\mathbf{I}_k$ is the $k \times k$ identity matrix. Finally, $\tilde{\boldsymbol{\Phi}}$ is another $Nk \times Nl$ block diagonal matrix with all the N diagonal entries set to $\boldsymbol{\Phi}$. Further, we'd like to explicitly point out that the matrix $\boldsymbol{\Sigma}_{\mathbf{Y}}{}^{-1}$ is a precision matrix and is not full-rank. Note that such non-invertible precision matrices are in-line with use of improper un-normalizable Gaussian distributions in Bayesian inference [9, 34] and do not pose any theoretical or practical problems.

The above mentioned marginalization preserves all the information about uncertainty due to the Gaussian Process regression part of the graphical model and is succinctly represented in the term $\boldsymbol{\Sigma}_{\mathbf{Y}}$. We propose to use variational inference and obtain an approximate posterior distribution over the label matrix $\mathbf{Y}$ that is a product of Gaussian terms and the corresponding Gamma terms over $\boldsymbol{\alpha}$. If we denote the approximation at the t^{th} iteration as $q^t(\mathbf{Y}) = N(\boldsymbol{\mu}_{\mathbf{Y}}^t, \boldsymbol{\Sigma}_{\mathbf{Y}}^t)$ and $q^t([\boldsymbol{\alpha}]) = \prod_i \Gamma(a_i^t, b_i^t)$, the resulting update rules are as following:

$$\text{Update for } q^{t+1}(\mathbf{Y}): \ \boldsymbol{\Sigma}_{\mathbf{Y}}^{t+1} = [\text{diag}(\mathbb{E}(\boldsymbol{\alpha}^t)) + \boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}]^{-1}$$

$$\text{Update for } q^{t+1}(\alpha^i): \ a_i^{t+1} = a_i^0 + 0.5,$$

$$b_i^{t+1} = b_i^0 + 0.5[\boldsymbol{\Sigma}^{t+1}(i,i)]$$

Given observed labels and $q(\mathbf{Y})$, the posterior distribution $q(\mathbf{Y}_{\mathcal{U}})$ over unobserved labels is simply the conditional Gaussian distribution obtained by using standard Gaussian identities. We'd like to point out that unlike the previous approach [17], the cross co-variances across all the labels for all the points are preserved in $\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}$. Further note that this algorithm results in the final variance over the label vectors in terms of the kernel function over the input features $\mathbf{X}$. This is a direct consequence of the fact that uncertainty from the regression part of the network has already been propagated to $\mathbf{Y}$ and sets a basis for an active learning scheme that is effective.

Selecting Hyperparameters a_0 and b_o: The choice of the hyperparameters a_0 and b_0 is critical to induce sparsity, as arbitrary assignments can lead to non-sparse solutions. To see this, consider the joint distribution $p(\mathbf{Y}, [\boldsymbol{\alpha}_i]_{i=1}^N | \mathbf{X}, \boldsymbol{\Phi})$ and analytically marginalize over $[\boldsymbol{\alpha}_i]_{i=1}^N$. This results in a distribution of the following form:

$$\int_{[\boldsymbol{\alpha}_i]_{i=1}^N} p(\mathbf{Y}, [\boldsymbol{\alpha}_i]_{i=1}^N | \mathbf{X}, \boldsymbol{\Phi}) = e^{-\frac{\mathbf{Y}^T \boldsymbol{\Sigma}_{\mathbf{Y}}^{-1} \mathbf{Y}}{2}} \prod_{i=1}^{N} \prod_{j=1}^{l} (1 + \frac{y_{ij}^2}{2b_0})^{-a_0}. \quad (4)$$

The above marginalization leads to a very intuitive interpretation where the term $\boldsymbol{\Sigma}_{\mathbf{Y}}{}^{-1}$ arises due to the regression part

[1]Extension to non-linear kernel is straightforward by replacing $\mathbf{X}^T\mathbf{X}$ with an appropriate kernel function.

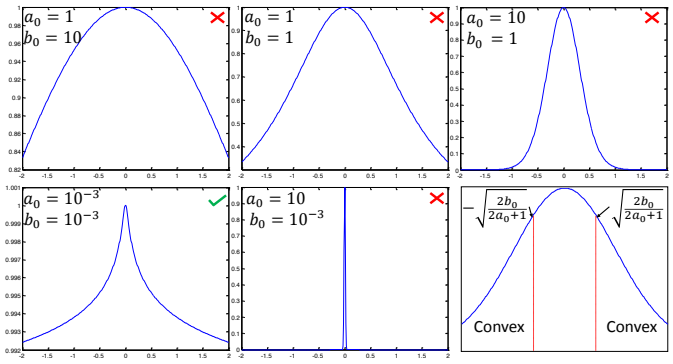


Figure 1: Plot of the sparsity inducing prior term $(1 + \frac{y^2}{2b_0})^{-a_0}$ for different values of a_0 and b_0. The desirable shape of the prior term is shown in bottom-left corner (indicated by a green ✓) and occurs when a_0 and b_0 are set to 10^{-3} . The bottom-right figure characterizes the shape as a function of a_0 and b_0 and highlights that the shape is undesirably concave when $-\sqrt{\frac{2b_0}{2a_0+1}} < y < \sqrt{\frac{2b_0}{2a_0+1}}$.

of the framework and the terms $(1+\frac{y_{ij}^2}{2b_0})^{-a_0}$ induce the sparsity in the solution. Note that in the sparsity term, higher values of y_{ij} lead to a lower value of the potential, thereby penalizing non-zero instantiations. However, the shape of the penalty term critically depends upon the values of a_0 and b_0. Figure 1 plots these penalty prior functions for different values of a_0 and b_0. Figure 1 shows several shapes corresponding to different hyperparameter settings. Most of the shapes shown in the figure are undesirable for inducing sparsity due to their non-convexity localized around a large range near zero (highlighted by a red ×). The bottom-right panel on the other hand shows what a desirable penalty function should look like ($a_0 = b_0 = 10^{-3}$) (denoted by a green ✓).

The bottom-right subpanel in figure 1 characterizes the shape of the function in terms of the hyperparameters. In particular, the function can decomposed into three regions: (1) $y < -\sqrt{\frac{2b_0}{2a_0+1}}$, (2) $y > \sqrt{\frac{2b_0}{2a_0+1}}$, and (3) $-\sqrt{\frac{2b_0}{2a_0+1}} < y < \sqrt{\frac{2b_0}{2a_0+1}}$. It is straightforward to show that the second derivative of the function will always be greater than zero for region (1) and (2), thereby implying a nice locally convex behavior leading to a desirable shape of the penalty function. The region (3) on the other hand has the second derivative less than zero, thus leading to a shape with a flatter penalty structure. Consequently, in order to induce an appropriate sparsity inducing prior it is desirable to minimize the range where region (3) occurs. This is achieved when either a_0 takes a very high value or when b_0 tends to zero. However, setting a_0 to a high value will result in a very peaky penalty function (see figure 1 bottom-middle panel). A peaky penalty function is undesirable as it will override the potential arising due to the regression term in the overall mutilabel framework. Thus, a desirable prior can only be achieved when we select a very small value for both a_0 and b_0. It is no surprise that in prior research on Bayesian compressed sensing [6, 17, 25], a_0 and b_0 were set to very small values. Our analysis above validates such choices in order to incorporate a reasonable sparsity inducing prior. This observation about requiring a_0 and b_0 to be close to zero will be critical in the next section where we discuss non-myopic active learning.

4.2 Non-Myopic Selective Sampling

The goal of active learning is to select a set of points $\mathcal{A}$ to label from the available pool of unlabeled data $\mathcal{U}$ under budget constraints, $|\mathcal{A}| = n$, such that the resulting classifier is most accurate. Intuitively, we want to sample the n most informative points with respect to the classification task. Two potential criteria for the task are entropy and mutual information. Selecting points with maximum entropy boils down to choosing a set of points that jointly have the maximum uncertainty on their labels. The mutual information criterion [7], on the other hand, chooses $\mathcal{A}$ so that the uncertainty over the labels of remaining points is maximally reduced after the labels of $\mathcal{A}$ are incorporated in the predictive model.

The mutual information based criterion is the best of the two heuristics as the entropy criterion tends to select points that are far apart from each other in the feature space and hence ends up selecting points on the boundary. Since a point usually provides information about points in its nearby region, selecting points on the boundary wastes information gathering effort. In this paper, we present results in the context of mutual information as a criterion. Formally, if $H(\cdot)$ denotes the entropy, then we write the non-myopic subset selection problem as:

$$\mathcal{A}^* = \underset{\mathcal{A}\subseteq\mathcal{U},|\mathcal{A}|=n}{\arg\max}\ H(\mathbf{Y}_{\mathcal{U}\setminus\mathcal{A}}) - H(\mathbf{Y}_{\mathcal{U}\setminus\mathcal{A}}|\mathbf{Y}_{\mathcal{A}}). \qquad (5)$$

We call this problem non-myopic due to the fact that the goal is to reason about the entire set $\mathcal{A}$ at once. This is different from the other schemes for active classification where a single point is chosen for querying a label, the model updated after observing the label and the process repeated for a total of n rounds. This non-myopic subset selection problem is NP-complete as shown below:

PROPOSITION 1. *The subset selection problem defined in equation 5 is NP-complete for the density $p(\mathbf{Y}, [\boldsymbol{\alpha}_i]_{i=1}^N|\mathbf{X}, \boldsymbol{\Phi})$.*

This proposition can easily be established by observing that, when the random variables $\boldsymbol{\alpha}$ are completely observed, the distribution reduces to the form of a multivariate Gaussian density. The proof simply follows from the fact that the subset selection problem for mutual information criterion has been previously shown to be NP-complete [19]. □

Given the hardness of this problem, we resort to a greedy approximation algorithm following recent ideas in submodular optimization. Specifically, if we can show that the objective (eq. 5) being optimized is submodular and non-decreasing, then we can derive a greedy strategy that sequentially selects data points based on marginal improvement of the objective. Note that the mutual information criterion is not submodular in general, but under some weak conditions, both submodularity and its non-decreasing property can be established. Formally, [18] have proved the following proposition for random variables $\mathcal{S}$ and $\mathcal{U}$ in a graphical model:

PROPOSITION 2. **[18]** *Let $\mathcal{S}$, $\mathcal{U}$ be disjoint subsets of random variables such that the variables in $\mathcal{S}$ are independent given $\mathcal{U}$. Let information gain be $F(\mathcal{A}) = H(\mathcal{U}) -$*

$H(\mathcal{U}\backslash\mathcal{A}|\mathcal{A})$, where $\mathcal{A} \subseteq \mathcal{U}$. Then F is submodular and non-decreasing on $\mathcal{U}$, and $F(\emptyset) = 0$.

The following observation identifies the graphical model corresponding to the joint distribution $p(\mathbf{Y}, [\boldsymbol{\alpha}_i]_{i=1}^N | \mathbf{X}, \boldsymbol{\Phi})$ as a special case of the above, thus establishing the submodularity and non-decreasing characteristic of our objective function.

COROLLARY 1. *Let $\mathcal{S} = [\boldsymbol{\alpha}]_{i=1}^n$ and $\mathcal{U} = \mathbf{Y}_\mathcal{U}$, then due to the form of $h_{\boldsymbol{\alpha}_i}$ and $p(\boldsymbol{\alpha}_i)$, we have $[\boldsymbol{\alpha}_i]_{i=1}^N$ independent given $\mathbf{Y}_\mathcal{U}$. Consequently, the objective in equation 5 is submodular and non-decreasing.* □

Given this observation, we can now propose a greedy algorithm that repeatedly picks the points with the highest increase in mutual information and adds them to the subset $\mathcal{A}$. The point $x^* \in \mathcal{U}$ selected to be added to the subset $\mathcal{A}$ is such that $x^* = \arg\max_x (MI(\mathcal{A} \cup x) - MI(\mathcal{A}))$. Here, $MI(\mathcal{A}) = H(\mathbf{Y}_{\mathcal{U}\backslash\mathcal{A}}) - H(\mathbf{Y}_{\mathcal{U}\backslash\mathcal{A}}|\mathbf{Y}_\mathcal{A})$ and it has been proved earlier that this algorithm selects the subset $\mathcal{A}^*$ such that the value of the objective function will at least be $(1 - \frac{1}{e})$ times the optimal solution [19, 22].

While the above mentioned corollary shows the existence of a greedy algorithm that has a good approximation guarantee, it is still non-trivial to compute the mutual information $MI(\mathcal{A})$ for the multilabel classification model. Specifically, in order to compute mutual information, we need to solve exact inference, which itself is non-trivial in our model. Here also, in order to derive an implementable solution, we need to establish a similar approximation $\hat{MI}$ to mutual information such that that the set of selected points $\mathcal{A}^*$ does not change when the approximated mutual information is used instead of the exact one. Formally, we prove the following theorem which in turn will imply that such an approximation is feasible:

Theorem 1. *Let $\hat{MI}$ denote the mutual information of any set $\mathcal{A}$ computed over the probability distribution, $p(\mathbf{Y}) \propto \exp[\mathbf{Y}^T \boldsymbol{\Sigma}_\mathbf{Y}^{-1} \mathbf{Y}/2]$, where $\boldsymbol{\Sigma}_\mathbf{Y}^{-1}$ is as defined in eq. 3, then for any $x \in \mathcal{U}$:*

$$\lim_{\substack{a_0 \to 0 \\ b_0 \to 0}} MI(\mathcal{A} \cup x) - MI(\mathcal{A}) = \hat{MI}(\mathcal{A} \cup x) - \hat{MI}(\mathcal{A}).$$

Proof Sketch: To prove this theorem, we first use the fact that $MI(\mathcal{A} \cup x) - MI(\mathcal{A})$ can be written as $H(x|\mathcal{A}) - H(x|\mathcal{U}\backslash x)$ [19]. Next, we consider the joint distribution over $\mathbf{Y}$ given in eq. (4) that arises after analytically marginalizing over $[\boldsymbol{\alpha}_i]_{i=1}^N$. It is easy to show that given $\mathcal{A}$, the conditional predictive distribution of x takes a similar form:

$$p(x|\mathcal{A}) \propto \exp[-\frac{(x - m_\mathcal{A})^2}{2\sigma_\mathcal{A}^2}](1 + \frac{x^2}{2b_0})^{-a_0}$$

$$\text{where: } m_\mathcal{A} = \boldsymbol{\Sigma}_{x,\mathcal{A}} \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}}^{-1} \mathbf{Y}_\mathcal{A}$$

$$\sigma_\mathcal{A}^2 = \boldsymbol{\Sigma}_{x,x} - \boldsymbol{\Sigma}_{x,\mathcal{A}} \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}}^{-1} \boldsymbol{\Sigma}_{\mathcal{A},x}$$

The above reduction follows from simple algebraic manipulation where the terms corresponding to the set $\mathcal{A}$ are collected by using matrix inversion lemma. A similar form also can be derived for the conditional predictive distribution of x given $\mathcal{U}\backslash x$. Let us denote the conditional entropy of x given $\mathcal{A}$ computed via the Gaussian distribution as $\hat{H}(x|\mathcal{A})$. The proof then follows from the fact that $\lim_{\substack{a_0 \to 0 \\ b_0 \to 0}} H(x|\mathcal{A}) = \hat{H}(x|\mathcal{A})$ (similarly for $H(x|\mathcal{U}\backslash x)$). Proof for the above statement is provided in appendix A. □

The implication of the above theorem is that the greedy selection strategy described above will yield nearly the same subset when $\hat{MI}$ is used instead of the true mutual information, as a_0 and b_0 are set to very small values in our algorithm. This is due to the fact that the greedy procedure seeks to include x such that $(MI(\mathcal{A} \cup x) - MI(\mathcal{A}))$ is maximized at each step and the above theorem guarantees that the order of selecting the points will not change when MI is replaced by $\hat{MI}$. This result enables us to derive an algorithm based on the approximate inference procedure described earlier in the paper. Note that the $\hat{MI}$ is defined over a Gaussian Process; thus, prior work [19] on near-optimal selective sampling can be directly applied here. Also, note that the discussion above only focuses where the Gaussian term has a zero mean for clarity purposes. It is fairly straightforward to extend the analysis to non-zero mean cases.

One of the surprising by-products of our result is the fact that, similar to Gaussian Process regression models, the observed labels do not affect the order in which the points should be selected, which in turn allows us to do non-myopic selection of points. Intuitively this is feasible by realizing that the mapping of the discrete label vectors to the continuous space via the matrix $\boldsymbol{\Phi}$ is almost 1-to-1 [15]. So, any subset selection via Gaussian Process regression on the continuous space, which is independent of observations, automatically transfers to the discrete labeled space. We wish to point out explicitly that $\boldsymbol{\Sigma}_\mathbf{Y}^{-1}$ is not full-rank; however, the computation over Gaussian mutual information is still feasible in such cases [14].

The proposed compressed sensing framework compresses the information about the labels, $\mathbf{Y}$ into the compressed space, $\mathbf{Z}$ and then couples the $\mathbf{Z}$ in a single covariance matrix, hence allowing us to do mutual information based active learning in time that is independent of the number of labels. The compressed sensing formulation also enables us to train lesser number of classifiers for the training process and reduce the training time significantly. Finally, note that the mutual information computation need not be done from scratch after each iteration of the active learner. We use lazy evaluations proposed in [19] to drastically reduce the computational cost for selection of subsequent points to be annotated.

5. ACTIVE LEARNING

It is interesting to note that selective sampling via $\hat{MI}$ can accommodate many different scenarios of selective sampling. The most popular setting in literature considers revealing all the labels for selected data points. Alternatively, in active diagnosis, different labels can be probed for one particular test case only. Finally, in the most general case, individual labels from different input points can be selected (generalized active learning).

5.1 Traditional Active Learning

In traditional active learning, we select a subset of points from the available data points for which all labels are revealed. Ideally, the goal is to choose a subset of size n that leads to maximum decrease in entropy over the remaining unlabeled points as described in equation 5. However, proposition 2 and corollary 1 allow us to greedily select points to be annotated so as to maximize the mutual information at each step. Algorithm 1 shows the outline of the method that allows efficient sampling of points based

Algorithm 1 Mutual information based active learning over the compressed space

Input: Input features $\mathcal{X}$ and budget n
Output: $\mathcal{A}$: The set of points to be labeled

```
Compute C = Σ_Z + χ^2 I using Σ_Z defined in eq. 3
A ← φ
for i ← 1 to n do
  for x ∈ X\A do
    δ_x ← |C(x,x) − C(x,A) C(A,A)^{-1} C(A,x)| / |C(x,x) − C(x,A) C(A,A)^{-1} C(A,x)|
  end
  x_* ← arg max_x δ_x
  A ← A ∪ x^*
end
```

on mutual information. Note that, for this case, when all the labels are revealed per data point, we can be far more efficient in computation by directly using the Kernel matrix defined over the points alone.

5.2 Active Diagnosis

Such an active learning scenario is more specific to multilabel active learning, where obtaining each label for a point has significant cost associated with it. For example, in industrial or medical settings, where each label may be obtained as a result of an expensive test, it is wiser to select the labels which you want to be annotated for a selected point to decrease the overall cost, as opposed to annotating all the labels. If we denote the label vector as $\mathbf{y} = \{y_1, y_2, y_3, ...y_l\}$, then the active diagnosis problem is to select a subset of labels, $l' \subset \{1, 2, ...l\}$, which need to be annotated. This scenario can be easily handled by our model as we obtain the complete covariance matrix over the labels and the mutual information criterion $\hat{MI}$ treats each label independently. We can follow a greedy algorithm similar to the one followed in algorithm 1 to sample labels, wherein at each step, we select the label with index, $i = \arg\max_{i \in \mathcal{U}} \hat{MI}(\mathbf{y}_{\mathcal{L} \cup i}) - \hat{MI}(\mathbf{y}_{\mathcal{L}})$ to be labeled, where $\mathcal{U}$ and $\mathcal{L}$ denote the unlabeled and labeled indices respectively, as before. We present evaluation of our scheme for active diagnosis in the results section.

5.3 Generalized Active Learning

A combination of the above active learning scenarios leads to a general active learning problem, wherein the goal is to select both the points and the labels to be annotated within a given active learning budget. Given the budget to obtain n annotations, we select n document-label pairs from the set of pairs $\{(1,1),(1,2),.....,(N,l)\}$, where N is the number of documents with l labels each. To achieve this objective, we consider the information gain computed using $\hat{MI}$ obtained over the complete label matrix $\mathbf{Y}$. Once again, we use the greedy strategy to pick each pair. However, computing this metric over the entire space of document-label pairs is expensive and we resort to a two-way approach, wherein we first select the document to be annotated based on algorithm 1 and then select, the label to be revealed similar to active diagnosis. Note that, both active diagnosis and generalized active learning are novel active learning scenarios which are specific to the multilabel classification problem and can't be handled by SVM based state-of-the-art methods.

6. RESULTS

In this section, we present empirical results to demonstrate a) how the proposed approximate inference procedure compares to the prior work [17](denoted as BML-CS) in terms of accuracy as well as the capability to estimate variances, b) the comparison of mutual information based sampling with the state-of-the-art SVM based active learning method proposed in [20], uncertainity sampling and random sampling baselines, and c) performance of the method on novel active learning scenarios like active diagnosis and generalized active learning which are specific to multilabel classification.

6.1 Setup

Datasets: We used the datasets listed in Table 1 to evaluate our algorithm. As can be seen from the table, the datasets exhibit a wide range in type, feature vector size and label vector size. For each of the datasets, 4000 randomly sampled points were used as the active learning pool from which points had to be selected for training based on the different active learning strategies compared. Another set of 2000 points was used as the test data. For datasets with fewer than 6000 points (`enron`, `medical`, `MSRC` and `yeast`), the entire set was selected as the active learning pool and testing was done on the points not selected for training. All the results are reported after averaging over 5 such splits. For all the methods, an initial seed of 500 randomly sampled points was provided to the algorithms to start with (50 points, in case of `MSRC`, and 200, in case of `enron`, `medical` and `yeast`), as has been done in standard active learning literature. For all our experiments, we used features downloaded from the Mulan multilabel datasets [1]. For `MSRC`, we used 1024 bit features generated using the Picodes scheme [4].

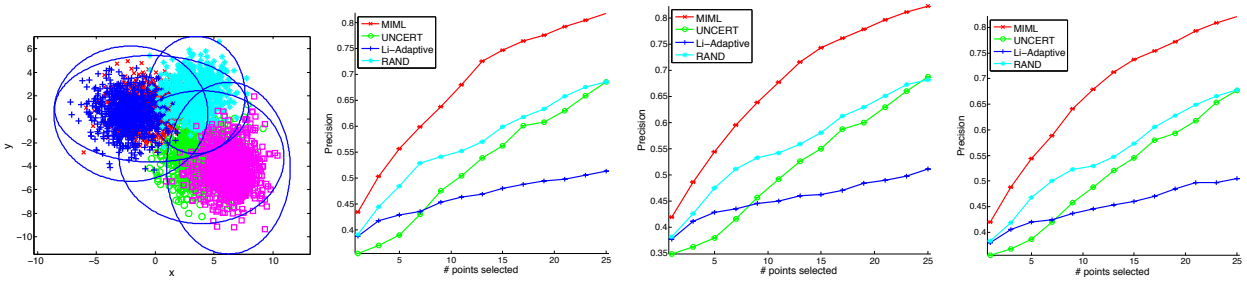
Table 1: Datasets Used for Experiments

Dataset	Type	Instances	Features	Labels
Yeast	Biology	2417	103	14
MSRC	Image	591	1024	23
Medical	Text	978	1449	45
Enron	Text	1702	1001	53
Mediamill	Video	43907	120	101
RCV1	Text	6000	47236	101
Bookmarks	Text	87856	2150	208
Delicious	Text	16105	500	983

Baselines We compare the following methods for the evaluation of the proposed active learning strategy:

- **MIML**: Mutual Information for Multilabel classification as proposed in this paper
- **Uncert**: Uncertainty sampling on the model proposed in this paper. This method picks the point with the highest entropy after each iteration.
- **Rand**: Random sampling baseline wherein each point is randomly sampled.
- **Li-Adaptive:** SVM based adaptive active learning method proposed in [20], which is the state-of-the-art in multilabel active learning.

Parameters: For MIML and Uncert, the dimension of the compressed label space k was set to half of the number



(a) Sample toy data, $L = 5$ (b) Precision at 1 (c) Precision at 2 (d) Precision at 3

Figure 2: Active Learning on Toy dataset:a)A sample toy dataset generated with 5 Gaussians. Points are generated from $L = 5$ Gaussians for this figure and are assigned labels based on the Euclidean distance from the Gaussian means. In our experiments, we use $L = 35$ Gaussians.b,c,d) The precision at 1,2 and 3 on the toy dataset obtained after averaging over 50 different runs. This clearly shows how mutual information quickly picks informative points to be annotated and hence, shows considerable increase in performance.

of labels for all datasets. The hyperparameters σ^2 and χ^2 in MIML and Uncert were selected via evidence maximization and a_0 and b_0 were set to 10^{-6}, which lead to fairly uninformative priors. The Li-Adaptive mechanism proposed in [20] has a hyperparameter β, which was selected from a prefixed set of values $0.1, 0.2, ..., 0.9, 1$ as suggested in [20]. Cross validation was done for all other parameters.

Performance Metric: Typically, the goal in multilabel classification is to predict the top k labels for each point. So, we use precision at 1, 2 and 3 to quantify the performance of different active learning schemes, which is in line with previous work. Precision at k is defined as the fraction of true positives over the top k predicted labels for each point. For novel active learning scenarios like generalized active learning and active diagnosis, the ultimate goal is to recover the whole label set of each single test point. So, we quantify our results using the F-1 measure which depends on all labels, rather than on the few top ones: F - measure $= 2\ \frac{\text{precision} * \text{recall}}{\text{precision}+\text{recall}}$.

6.2 Toy Dataset

We first consider a toy dataset before assessing the performance of the different schemes on real-world datasets in 6.4. The toy dataset is created as follows: We create $L = 35$ two-dimensional Gaussians and assign a unique identifier to each of them. Their mean and standard deviation is randomly selected from a fixed interval such that the contours of the Gaussians, drawn for C standard deviation away from the mean, overlap with each other. Points are generated from each of the L Gaussians. The feature vector is just the 2-dimensional coordinate of the point and the labels are the Gaussians whose mean is no more than C standard deviations away from the point. For the experiments, C was selected such that each point belongs to 7 classes on average. To visualise the dataset, we plot the dataset generated by setting $L = 5$, i.e. a dataset generated with 5 Gaussians in Figure 2(a). For our experiments, we use $L = 35$ and start with a randomly sampled set of 20 points. We, then, select 25 more points based on different schemes from a pool of 1000 data points. The performance is averaged over 50 different runs and compared across all methods mentioned in section 6.1.

The precision at 1, precision at 2 and precision at 3 is shown in figure 2. As can be seen in the figure, the mutual information criterion quickly fills the training set with points that are helpful to the model. Uncertainty sampling on our model outperforms Li-Adaptive as points are added. In fact, Li-adaptive baseline performs worse than the random sampling baseline in this case. Intuitively, MIML starts by picking points across different Gaussians in order to maximize mutual information and hence leads to better prediction. On the other hand, the uncertainty sampling criterion picks points closer to the boundaries, hence missing out on information.

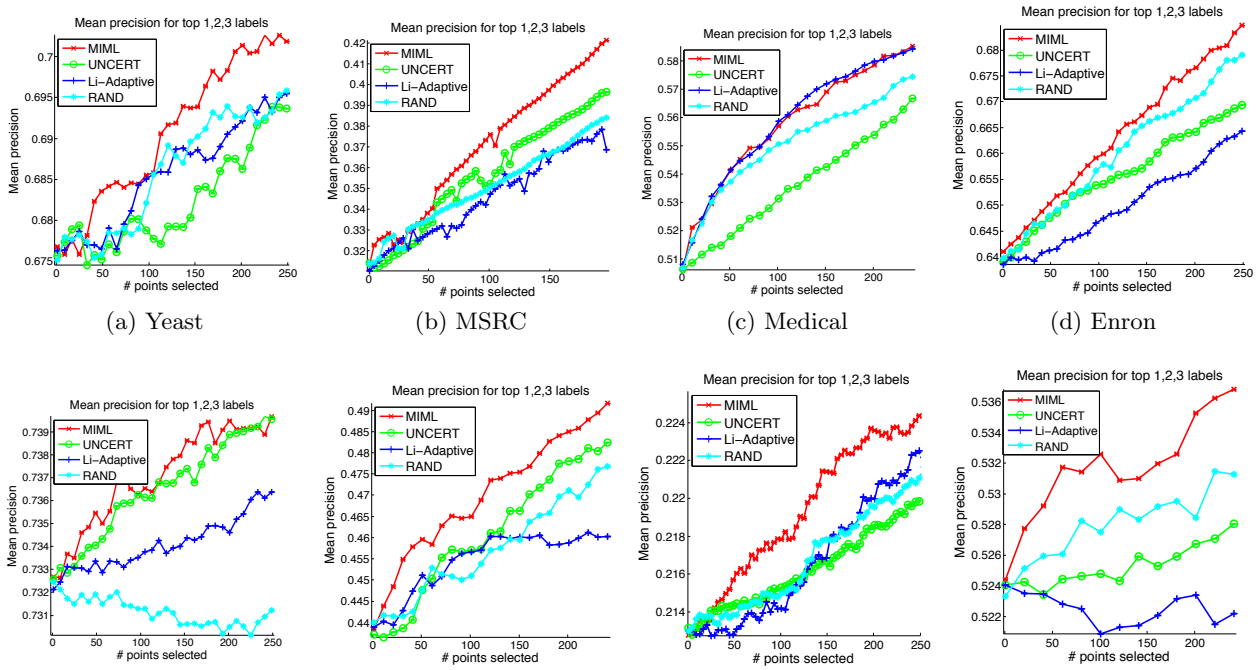
6.3 Exploration of the Alternate Inference Procedure:

In order to explore the properties of inference procedures (BML-CS [17] and proposed method(ML-OSS)), we compare the performance of both algorithms on the `MSRC` dataset, which has 23 labels. We used 1024-bit Picodes [4] image descriptors as features.

The results of the experiments are shown in Figure 4(a) and 4(b). Figure 4(a) plots the means of the labels inferred by both the methods for five randomly sampled test points and highlights that the means inferred by both the methods are highly correlated indicating that they are equally capable of modeling the means of the posterior distribution. However, similar correlation is not observed for variance estimates. Figure 4(b) plots the variances for the first label of all the test images inferred by the proposed method vs BML-CS on a negative log scale. Variances inferred by our inference method have a much wider spread indicating that BML-CS underestimates variances due to its limitation in propagating variances across the graphical model.

6.4 Evaluation of Active Learning Strategy

We compared the performance of the proposed mutual information (MIML) strategy with the strategy proposed in [20], uncertainty sampling (UNCERT) and random sampling (Rand). In these experiments, at every active learning round, all the labels corresponding to an input data point were revealed. Figure 3 shows the average of precision at 1,2 and 3 achieved on different datasets. Note that this is a significant metric as all our datasets except `delicious` have less than 3 positive labels per document on average.



(a) Yeast (b) MSRC (c) Medical (d) Enron

(e) Mediamill (f) RCV1 (g) Bookmarks (h) Delicious

Figure 3: Active Learning with Mutual Information (IG): The averages of precision at 1, 2 and 3 on different datasets as points get added to the training set. Our method, MIML, comprehensively outperforms uncertainty sampling(UNCERT) and the random sampling baseline(Rand) as well as the state-of-the-art SVM based active learning approach, Li-Adaptive [20]. None of the other methods shows such consistent performance across the diverse range of datasets. Every other method is outperformed by the random sampling baseline in one or the other dataset.

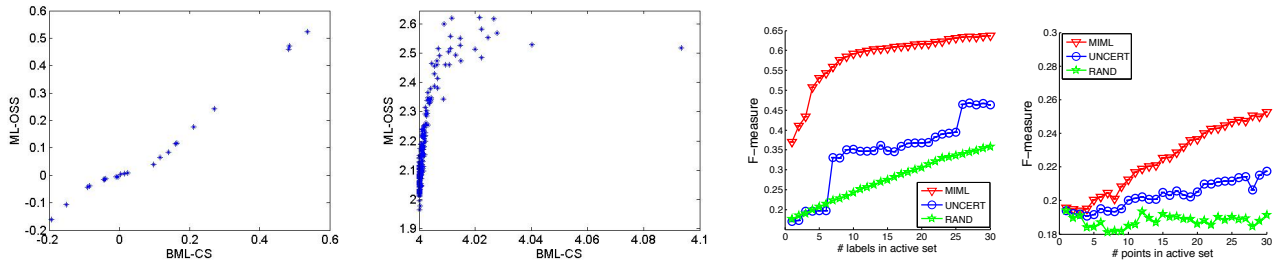
Dataset	Labels	MIML	Li-Adaptive
Yeast	14	3m 25s	1m 54s
Mediamill	101	41m 29s	54m 35s
RCV1	101	30m 45s	37m 35s
Bookmarks	208	48m 58s	3h 57m
Delicious	983	1h 11m	20h 15m

Table 2: Time complexity analysis: Time taken to select 250 points from a pool of 4000 points for different datasets for our method (MIML) and the state-of-the-art SVM based approach [20]. For Yeast, the pool is 2000 points and other smaller datasets have been excluded. MIML scales much better than Li-Adaptive with the number of labels and achieves nearly 20x gain for Delicious.

We observe that the MIML criterion proposed in this paper outperforms all the compared methods consistently, on all the datasets. No other method beats the other methods across all datasets. For instance, Li-adaptive performs worse than random sampling on almost half of the datasets.

Evaluation of time complexity: Table 2 shows the time taken by the proposed method (MIML) and Li-adaptive, to select 250 points to be annotated from a pool of 4000 points. We do not include smaller datasets in this analysis. For `Yeast`, the pool is 2000 points as the dataset has only 2417 points. All these experiments were performed on a standard desktop PC with a 4 core, hyper-threaded Intel Core-i7 3.4GHz processor and 16GB RAM. As can be seen, the SVM based approach scales badly (even though training of binary classifiers per label was parallelized) with the number of labels and takes approximately 20 times the time taken by MIML on `delicious`. The only overhead in MIML with increasing number of labels is because of the time taken to learn the Gaussian process over the labels with an initial set of points. All other updates can be performed incrementally and hence, MIML is extremely time-efficient.

Active Diagnosis and Generalized Active Learning: Next, we demonstrate two active learning strategies specific to multilabel classification and show the benefits of our mutual information based strategy in these scenarios. As described before, in active diagnosis, one label per round is revealed for the test set based on the selection criteria. This scenario cannot be straightforwardly tackled by the method of [20], whereas it naturally fits in our model, as the employed joint distribution correlates both points and labels. Thus, we compare our results against the UNCERT and random sampling baselines. Since the ultimate goal is to recover the whole label set of each single test point, we quantify our results using the F-1 measure which depends on all labels, rather than on the few top ones: F - measure $= 2\ \frac{\text{precision} * \text{recall}}{\text{precision}+\text{recall}}$. We report the results of this experiment on the `RCV1` dataset; we randomly selected $n_* = 30$ points collected in a test set $\mathbf{Y}_*$ for which we did active diagnosis by selectively sampling their labels. Each test point has an unknown label vector of size $l = 101$ and the active diagnosis procedure was given a budget of



(a) Comparing means (b) Comparing Variances (c) Active Diagnosis (d) Generalized AL

Figure 4: (a, b) Comparison of the inference procedure with BML-CS: Figure (a) is a plot of means propagated by the proposed method (ML-OSS) vs means propagated by BML-CS for 5 randomly sampled test points. Mean values propagated by both the methods follow similar trends. Figure (b) plots variances inferred over the test points by our method vs variances propagated by BML-CS. Both the axis have been converted to negative log scale. Variances propagated by our approach have a wider spread and are higher. (c) Active Diagnosis: Plot of mean F-score against the number of active diagnosis rounds. Mutual information (MIML) outperforms uncertainity sampling (UNCERT) and random sampling (Rand). (d) Generalized Active Learning: Plot of mean F-score against the number of revealed points. Mutual information proves advantageous over the other methods.

$m = 30$ labels to select per point. An initial fully observed training set of 100 points kick started the experiment.

Fig. 4(c) summarizes the results in a plot averaged over 35 runs. We observe that the proposed criterion selects much more useful labels to reveal right from the start. The difference in the performance compared to the baselines is even larger than for the traditional active learning task. This is due to the fact that output labels within a data point are far more correlated than across data instances. For example, a test data corresponding to an article tagged as "investment" is very likely to also fall into the category "banking".

Generalized Active Learning is the scenario when both the input point and the label to be queried are selected based on the selection criteria. Since the goal here is to learn as much information as possible for the whole test set, we use the following criterion to evaluate the performance: after selecting $m = 30$ labels for each of the chosen data points, we compute the F-measure of the *whole* test set given the partially revealed label vectors of the current and all other data points already in the active set. We ran this experiment 35 times and present the averaged results for F-measure with number of points in the active set in figure 4(d). Note that, both active diagnosis and generalized active learning are enabled by the fact that our multilabel classification model can train on partially labelled datasets.

7. CONCLUSION

We presented a novel mutual information based active learning framework for multilabel classification, that enables a theoretically principled and non-myopic approach. We extensively evaluated our algorithm across various datasets in traditional active learning settings as well as active diagnosis and generalized active learning (which are specific to multilabel classification) and showed that it consistently outperforms the state-of-the-art, both in terms of time-efficiency and precision. Possible future work includes integration of the framework with other multilabel classification techniques.

8. REFERENCES

[1] Mulan Multilabel Datasets. http://mulan.sourceforge.net/datasets.html.

[2] R. Agrawal, A. Gupta, Y. Prabhu, and M. Varma. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *WWW*, 2013.

[3] K. Balasubramanian and G. Lebanon. The Landmark Selection Method for Multiple Output Prediction. In *ICML*, 2012.

[4] A. Bergamo, L. Torresani, and A. W. Fitzgibbon. PiCoDes: Learning a Compact Code for Novel-Category Recognition. In *NIPS*, 2011.

[5] W. Bi and J. T.-Y. Kwok. Efficient Multi-label Classification with Many Labels. In *ICML*, pages 405–413, 2013.

[6] C. M. Bishop and M. E. Tipping. Variational Relevance Vector Machines. In *UAI*, 2000.

[7] W. Caselton and J. Zidek. Optimal monitoring network designs . *Statistics and Probability Letters*, 1984.

[8] Y.-N. Chen and H.-T. Lin. Feature-aware Label Space Dimension Reduction for Multi-label Classification. In *NIPS*, pages 1538–1546, 2012.

[9] W. Chu, V. Sindhwani, Z. Ghahramani, and S. Keerthi. Relational Learning with Gaussian Processes. In *NIPS*, 2006.

[10] M. Cissé, N. Usunier, T. Artières, and P. Gallinari. Robust Bloom Filters for Large MultiLabel Classification Tasks. In *NIPS*, pages 1851–1859, 2013.

[11] A. Esuli and F. Sebastiani. Active Learning Strategies for Multi-Label Text Classification. In *ECIR*, 2009.

[12] C.-S. Feng and H.-T. Lin. Multi-label Classification with Error-Correcting Codes. *JMLR*, pages 289–295, 2011.

[13] A. Goldberg, X. Zhu, A. Furger, and J. Xu. OASIS: Online Active Semi-Supervised Learning. In *AAAI*, 2011.

[14] A. Gretton, R. Herbrich, and A. Hyvärinen. Kernel methods for measuring independence. *JMLR*, 2005.

[15] D. Hsu, S. Kakade, J. Langford, and T. Zhang. Multi-Label Prediction via Compressed Sensing. In *NIPS*, 2009.

[16] S. Ji, L. Tang, S. Yu, and J. Ye. Extracting Shared Subspace for Multi-label Classification. In *KDD*, pages 381–389, 2008.

[17] A. Kapoor, R. Viswanathan, and P. Jain. Multilabel Classification using Bayesian Compressed Sensing. In *NIPS*, 2012.

[18] A. Krause and C. Guestrin. Near-optimal Nonmyopic Value of Information in Graphical Models. In *UAI*, 2005.

[19] A. Krause, A. Singh, and C. Guestrin. Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies. *JMLR*, 2008.

[20] X. Li and Y. Guo. Active Learning with Multi-label SVM Classification. In *IJCAI*, 2013.

[21] X. Li, L. Wang, and E. Sung. Multi-label SVM Active Learning for Image Classification. In *ICIP*, 2004.

[22] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 1978.

[23] J. C. Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances In Large Margin Classifiers*. MIT Press, 1999.

[24] B. Settles. Active learning literature survey. Technical report, 2010.

[25] Shihao, Y. Xue, and L. Carin. Bayesian Compressive Sensing , 2007.

[26] A. Singh, A. Krause, C. Guestrin, and W. J. Kaiser. Efficient Informative Sensing using Multiple Robots.

[27] M. Singh, E. Curran, and P. Cunningham. Active learning for multi-label image annotation. Technical report, University College Dublin, 2009.

[28] F. Tai and H.-T. Lin. Multi-label Classification with Principal Label Space Transformation. In *Workshop proceedings of learning from multi-label data*, 2010.

[29] J. Weston, S. Bengio, and N. Usunier. Wsabie: Scaling Up To Large Vocabulary Image Annotation. In *IJCAI*, 2011.

[30] J. Weston, A. Makadia, and H. Yee. Label Partitioning for Sublinear Ranking. In *ICML*, 2013.

[31] B. Yang, J. Sun, T. Wang, and Z. Chen. Effective multi-label active learning for text classification. In *KDD*, 2009.

[32] H.-F. Yu, P. Jain, and I. S. Dhillon. Large-scale Multi-label Learning with Missing Labels. *ICML*, 2014.

[33] Y. Zhang and J. G. Schneider. Multi-Label Output Codes using Canonical Correlation Analysis. In *AISTATS*, pages 873–882, 2011.

[34] X. Zhu, J. Lafferty, and Z. Ghahramani. Semi-Supervised Learning: From Gaussian Fields to Gaussian Processes. Technical report, School of CS, CMU, 2003.

APPENDIX

Appendix A: Proof $\lim_{\substack{a_0\to 0\\ b_0\to 0}} H(x|\mathcal{A}) = \hat{H}(x|\mathcal{A})$

By definition $H(x|\mathcal{A}) = E_{\mathbf{Y}_\mathcal{A}}[H(x|\mathbf{Y}_\mathcal{A})]$ (also for $\hat{H}$). Here $E_{\mathbf{Y}_\mathcal{A}}[\cdot]$ is the expectation over the labels $\mathbf{Y}_\mathcal{A}$ at the active sites $\mathcal{A}$ under the distribution in eq. 4. Now,

$$\hat{H}(x|\mathbf{Y}_\mathcal{A}) - H(x|\mathbf{Y}_\mathcal{A}) = \frac{1}{\hat{Z}}\int_x e^{\frac{-(x-m_\mathcal{A})^2}{2\sigma^2_\mathcal{A}}}\frac{(x-m_\mathcal{A})^2}{2\sigma^2_\mathcal{A}}$$

$$-\frac{1}{Z}\int_x \frac{e^{\frac{-(x-m_\mathcal{A})^2}{2\sigma^2_\mathcal{A}}}}{[1+\frac{x^2}{2b_0}]^{a_0}}[\frac{(x-m_\mathcal{A})^2}{2\sigma^2_\mathcal{A}} + a_0\log(1+\frac{x^2}{2b_0})] + \log\frac{\hat{Z}}{Z}$$

Here, Z and $\hat{Z}$ are the corresponding normalizing constants. Now as $a_0 \to 0$, the term $a_0\log(1+\frac{x^2}{2b_0})$ vanishes. Further, using the binomial expansion $[1+\frac{x^2}{2b_0}]^{-a_0} = 1 - a_0\cdot t + \frac{a_0(a_0-1)}{2!}\cdot t^2 + ..$, where $t = \frac{x^2}{2b_0+x^2}$, it is straightforward to show:

$$\lim_{\substack{a_0\to 0\\ b_0\to 0}} \hat{H}(x|\mathbf{Y}_\mathcal{A}) - H(x|\mathbf{Y}_\mathcal{A}) = \hat{H}(x|\mathbf{Y}_\mathcal{A})[1-\frac{\hat{Z}}{Z}] + \log\frac{\hat{Z}}{Z}$$

The required proof follows directly by using the binomial expansion and seeing that as $a_0 \to 0$, the expression $\frac{\hat{Z}}{Z} \to 1$ and the above quantity evaluates to zero.