ANALYZING AVIATION SAFETY REPORTS: FROM TOPIC MODELING TO SCALABLE MULTI-LABEL CLASSIFICATION

AMRUDIN AGOVIC*, HANHUAI SHAN*, AND ARINDAM BANERJEE*

Abstract. The Aviation Safety Reporting System (ASRS) is used to collect voluntarily submitted aviation safety reports from pilots, controllers and others. As such it is particularly useful in researching aviation safety deficiencies. In this paper we address two challenges related to the analysis of ASRS data: (1) the unsupervised extraction of meaningful and interpretable topics from ASRS reports and (2) multi-label classification of ASRS data based on a set of predefined categories. For topic modeling we investigate the practical usefulness of Latent Dirichlet Allocation (LDA) when it comes to modeling ASRS reports in terms of interpretable topics. We also utilize LDA to generate a more compact representation of ASRS reports to be used in multi-label classification. For multi-label classification we propose a novel and highly scalable multi-label classification algorithm based on multi-variate regression. Empirical results indicate that our approach is superior to several baseline and state-of-the-art approaches.

1. Introduction

The Aviation Safety Reporting System (ASRS) [1] is used to collect voluntarily submitted aviation safety reports from pilots, controllers and others. The ASRS database is rich and constantly increasing in size. An ASRS report corresponding to a flight includes certain categorical values along with a text description. Each report is manually categorized and may belong to several categories simultaneously such as "maintenance problems" or "weather problems." The analysis of the data within the ASRS database plays an important role in furthering aviation safety, as it can be used to identify deficiencies and research human performance errors among other things.

In this paper we address two important hurdles one faces when analyzing the ASRS data. The first hurdle is to infer the key problems that are being discussed across different reports. When researching a specific kind of problem, one might be interested in knowing whether there are other reports dealing with a similar issue. Unfortunately manually defined categories alone might not be sufficient for this purpose. Such categories may be too high-level or coarse-grained, e.g., "maintenance problem" may refer to several rather different problems. Further, reports might discuss problems shared across multiple different pre-defined categories. Similarly there may be several subgroups of issues within a given category. In some cases, the manual categorization of reports may even be incorrect. Being able to analyze the data in terms of the underlying topics is therefore crucial. The second hurdle concerns automatically labeling the reports according to the pre-defined categories based on its topics of discussion. The key challenge stems from the fact that the problem is not one of standard classification since a report can have multiple labels simultaneously. Further, there may be correlations among the pre-defined categories which need to be taken into account while generating a multi-label prediction. Finally, the methods should be highly scalable in order to efficiently learn and male predictions on tens- or hundreds of thousands of reports and hundreds of classes.

We propose to use latent Dirichlet allocation (LDA), an existing state-of-the-art topic modeling approach, to automatically extract topics which are being discussed across ASRS reports. LDA is a hierarchical mixture model where each document is represented as a mixture of topics, and each topic is modeled as a distribution over words. We wish to investigate to what extent this model

*Department of Computer Science and Engineering, University of Minnesota, Twin Cities, aagovic@cs.umn.edu, shan@cs.umn.edu, banerjee@cs.umn.edu.

could be used on the ASRS data to extract meaningful and interpretable topics. In addition to analyzing underlying topics we utilize LDA to generate a lower-dimensional feature representation which we subsequently use in our classification task.

To address the problem of multi-label classification we propose Bayesian Multivariate Regression (BMR), a novel and highly scalable algorithm for multi-label classification. Our approach was designed to handle several challenges within the ASRS data. Each document in ASRS database is usually assigned to multiple categories, since there might be multiple problems occurring within the same flight. The categories (problems) are usually correlated. For instance, the "weather problem" tends to be correlated with the "landing problem", since bad weather increases the difficulty of landing. The conventional strategy of decomposing the multi-label prediction problem to multiple independent binary classification problems does not work well in this setting. Another challenge with the ASRS data is its sheer size. A multi-label classification algorithm in this setting needs to be both effective and highly scalable. Unlike most existing methods, BMR is capable of capturing correlations among classes, while being readily scalable to very large datasets. These are desirable properties which are useful beyond the domain of aviation safety. We compare our approach to two state-of-the-art methods and two one-versus-rest approaches. Our experimental results indicate superior performance across all used evaluation measures.

Overall the main focus of this work is the analysis of the ASRS data. Our contribution consists of two parts. The first part is applied in the sense that we investigate the usability of an existing topic model in the context of ASRS. The second part, the development of a multi-label classification, is an entirely novel contribution.

The rest of the paper is organized as follows: In Section 2, we give a brief overview on related work, including the topic modeling algorithms and multi-label classification algorithms. In Section 3, we propose our Bayesian Multivariate Regression approach and a variational inference algorithm to learn the model. We present the experimental results on ASRS dataset in Section 4, and conclude in Section 5.

2. Related Work

In this section we give a brief overview of existing topic modeling algorithms such as Latent Dirichlet Allocation [6] as well as several multi-label classification algorithms.

2.1. **Topic models.** Latent Dirichlet allocation (LDA) [6] is one of the most widely used topic modeling algorithms. It is capable of extracting topics from documents in an unsupervised fashion. In LDA, each document is assumed to be a mixture of topics, whereby a topic is defined to be a distribution over words. LDA assumes that each word in a document is drawn from a topic z, which in turn is generated from a discrete distribution Discrete(π) over topics. Each document is assumed to have its own distribution Discrete(π), whereby all documents share a common Dirichlet prior α. The graphical model of LDA is in Figure 1, and the generative process for each document $\mathbf{w}$ is as follows:

(1) Draw $\pi \sim$ Dirichlet(α).
(2) For each of m words $(w_j, [j]_1^m)$ in $\mathbf{w}$:
    (a) Draw a topic $z_j \sim$ Discrete(π).
    (b) Draw w_j from $p(w_j|\beta, z_j)$.

where $\beta = \{\beta_i, [i]_1^k\}$ is a collection of parameters for k topic distributions over totally V words in the dictionary. The generative process chooses β_i corresponding to z_j. The chosen topic distribution β_i is subsequently used to generate the word w_j. The most likely words in β_i are used as a representation for topic i.

Other than LDA, recent years have seen a large amount of work on topic modeling. Some examples include correlated topic models [3], dynamic topic models [4], and supervised topic models [5]. Correlated topic models capture the correlation among topics, while dynamic topic models capture the evolution of topics over time. Supervised topic models incorporate an additional response variable

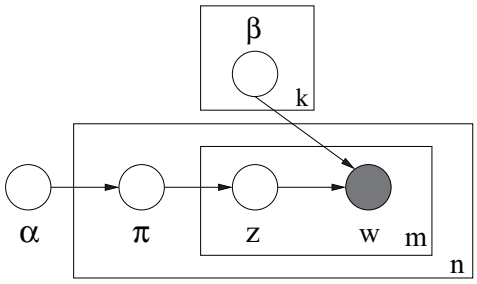


FIGURE 1. Graphical model for Latent Dirichlet Allocation.

into the topic model. For our purposes we chose to use LDA, because it is the least complex, and it is known to work well. Also note, as the size of the data set increases, the effect of assumed priors is minimized. In our case, the ASRS dataset is rather large.

2.2. **Multi-label classification algorithms.** Conventionally, multi-label classification problems were solved by decomposing them into multiple independent binary classification problems, while ignoring relationships between labels. In recent years, several approaches have been proposed which attempt to utilize the correlation structure among labels.

Kernel methods for multi-label classification tend to be extensions of the maximum margin idea. In [9], a maximum margin approach is proposed which minimizes the ranking loss. In [16], a method is proposed to learn a kernel which is shared across labels, to be subsequently used in individual label classifiers. While the ability to handle kernels is important in several domains, most existing approaches do not have a natural way of dealing with missing labels and are not probabilistic, i.e., no direct uncertainty quantification.

A number of probabilistic models have also been proposed for multi-label classification. In [12], a mixture model is proposed for text classification. More recently, in [13], a fully Bayesian model was proposed based on sparse and infinite canonical correlation analysis. It directly models correlations among labels and is one of few models which has the flexibility of dealing with missing labels. An extension of Gaussian Process prediction to the multi-label setting was proposed in [15].

The state-of-the-art also includes two approaches based on the k-nearest neighbor idea. In [17], label statistics from neighborhoods are used to build a Bayesian classifier. In [8], features are constructed based on label information from neighborhoods and subsequently used in logistic regression. In recent years, a family of methods based on multi-label dimensionality reduction has emerged [18, 10]. Our proposed model also falls in this category. Another interesting approach is presented in [7], where semi-supervised multi-label classification is proposed using the Sylvester equation.

There are two major problems with most existing approaches. They have a tendency not to explicitly model correlations among labels, but rather attempt to indirectly incorporate them. The second issue is that most existing approaches are too complex to be applicable to large scale datasets. Unlike most existing methods, our approach is a scalable probabilistic method which explicitly models the correlation structure among labels.

3. Bayesian Multivariate Regression

In multi-label classification, every data object is associated with a subset of possible labels. Assuming a total of c possible labels $L = \{\ell_1, \ldots, \ell_c\}$, for any given data object $\mathbf{x}$, the label information can be captured by a c-length bit vector $\mathbf{h} \in \{0, 1\}^c$, where $h_s = 1$ denotes the membership of $\mathbf{x}$ in class s.

3.1. **The Model.** We now introduce our novel approach which we call Bayesian Multivariate Regression (BMR). For simplicity we transform our binary labels h_s to truncated log odds $y_s \in \{-C, C\}$,

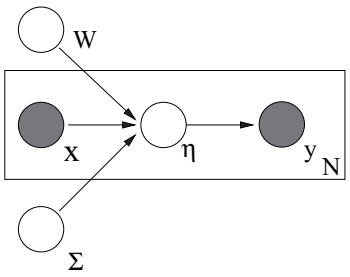


FIGURE 2. Graphical model for Bayesian Multivariate Regression.

where $C \in \mathbb{R}$. Log odds are defined as $\log\{p(h_s = 1)/(1 - p(h_s = 1)\}$, for binary labels these values are in $\{-\infty, +\infty\}$. By truncating the log odds we are effectively performing a relaxation of the problem. Rather than modeling binary vectors directly, our approach thus performs multivariate regression over the corresponding truncated log odds. Given a real valued feature vector $\mathbf{x} \in \mathbb{R}^k$ we assume a mapping $W \in \mathbb{R}^{c \times k}$, such that $\mu(\mathbf{x}) = W\mathbf{x}$. Subsequently we draw a latent label vector representation $\boldsymbol{\eta}$ from $N(\mu(\mathbf{x}), \Sigma)$, where $\Sigma \in R^{c \times c}$ denotes a covariance matrix among classes. While the covariance Σ is global in our model, the mean $\mu(\mathbf{x})$ differs for every data point. Our latent variable can alternatively be expressed as

$$\boldsymbol{\eta} = W\mathbf{x} + \zeta$$

where $\zeta \sim N(0, \Sigma)$. From this we can see that the empirical covariance of $\boldsymbol{\eta}$ will not be solely determined by Σ, but rather jointly by the mean function $\mu(\mathbf{x})$ and Σ. The last step in our model is to sample the label vector $\mathbf{y}$ from $N(\boldsymbol{\eta}, I)$. Integrating out the latent variable $\boldsymbol{\eta}$, allows us to incorporate the effects of Σ into W. Since it does not consider the marginal distribution over $\mathbf{x}$, BMR is a discriminative model.

Let $\mathbf{x}_n$ denote a k-dimensional data point, the generative process for each label c-dimensional label vector $\mathbf{y}_n$ can be specified as follows:

(1) $\boldsymbol{\eta}_n \sim N(W\mathbf{x}_n, \Sigma)$.
(2) $\mathbf{y}_n \sim N(\boldsymbol{\eta}_n, I)$.

The graphical model for BMR is shown in Figure 2. Given the model, the likelihood function of $\mathbf{y}_n$ is given by

$$\begin{aligned} p(\mathbf{y}_n|\mathbf{x}_n, \Sigma, W) &= \int_{\boldsymbol{\eta}_n} p(\boldsymbol{\eta}_n, \mathbf{y}_n|\mathbf{x}_n, \Sigma, W) d\boldsymbol{\eta}_n \\ &= \int_{\boldsymbol{\eta}_n} p(\boldsymbol{\eta}_n|W\mathbf{x}_n, \Sigma) p(\mathbf{y}_n|\boldsymbol{\eta}_n) d\boldsymbol{\eta}_n \ . \\ &= E_{\boldsymbol{\eta_n}}[p(\mathbf{y}|\boldsymbol{\eta_n})] \end{aligned} \tag{1}$$

Therefore, for a dataset with N data points $X = \{\mathbf{x}_n, [n]_1^N\}$ ($[n]_1^N \equiv n = 1 \ldots N$) and $Y = \{\mathbf{y}_n, [n]_1^N\}$, the likelihood function is

$$\begin{aligned} p(Y|X, \Sigma, W) &= \prod_{n=1}^{N} \int_{\boldsymbol{\eta}_n} p(\boldsymbol{\eta}_n|W\mathbf{x}_n, \Sigma) p(\mathbf{y}_n|\boldsymbol{\eta}_n) d\boldsymbol{\eta}_n \ . \\ &= \prod_{n=1}^{N} E_{\boldsymbol{\eta_n}}[p(\mathbf{y}|\boldsymbol{\eta_n})] \ . \end{aligned} \tag{2}$$

3.2. **Inference and learning.** For given data points X and corresponding Y, the learning task of BMR involves finding the model parameters W and Σ, such that the likelihood of $p(Y|X, \Sigma, W)$ as in Equation (2) is maximized. A general approach for such a task is to use multivariate optimization algorithms. However, the likelihood function in (2) is intractable, implying that a direct application

of optimization is infeasible. Therefore, we propose a variational inference method, which alternates between obtaining a tractable lower bound to the true log-likelihood and choosing the model parameters W and Σ to maximize the lower bound.

In order to obtain a tractable lower bound to (1), instead of using the true latent variable distribution $p(\boldsymbol{\eta}_n|W\mathbf{x}_n, \Sigma)$ in expectation calculation, we introduce a family of parameterized variational distributions $q(\boldsymbol{\eta}_n|\hat{\boldsymbol{\mu}}_n, \hat{\Sigma}_n)$ as an approximation to $p(\boldsymbol{\eta}_n|W\mathbf{x}_n, \Sigma)$, where $q(\boldsymbol{\eta}_n|\hat{\boldsymbol{\mu}}_n, \hat{\Sigma}_n)$ is a Gaussian distribution, and $\hat{\boldsymbol{\mu}}_n$ and $\hat{\Sigma}_n$ are variational parameters denoting the mean and covariance. Following Jensen's Inequality [6], we have

$$\begin{aligned}(3)\quad \log p(\mathbf{y}_n|\mathbf{x}_n, \Sigma, W) &\geq E_q[\log p(\boldsymbol{\eta}_n, \mathbf{y}_n|\mathbf{x}_n, W, \Sigma)] - E_q[\log q(\boldsymbol{\eta}_n|\hat{\boldsymbol{\mu}}_n, \hat{\Sigma}_n)] \\ &= E_q[\log p(\boldsymbol{\eta}_n|\mathbf{x}_n, W, \Sigma)] + E_q[\log p(\mathbf{y}_n|\boldsymbol{\eta}_n)] - E_q[\log q(\boldsymbol{\eta}_n|\hat{\boldsymbol{\mu}}_n, \hat{\Sigma}_n)] \ .\end{aligned}$$

We can denote the lower bound (3) using $L(\hat{\boldsymbol{\mu}}_n, \hat{\Sigma}_n, W, \Sigma)$, and each term in $L(\hat{\boldsymbol{\mu}}_n, \hat{\Sigma}_n, W, \Sigma)$ are given by

$$\begin{aligned}E_q[\log p(\boldsymbol{\eta}_n|\mathbf{x}_n, W, \Sigma)] &= -\frac{1}{2}\left(\mathrm{Tr}(\Sigma^{-1}\hat{\Sigma}_n) + (\hat{\boldsymbol{\mu}}_n - W\mathbf{x}_n)^T\Sigma^{-1}(\hat{\boldsymbol{\mu}}_n - W\mathbf{x}_n)\right) - \frac{c}{2}\log 2\pi + \frac{1}{2}\log|\Sigma^{-1}| \\ E_q[\log p(\mathbf{y}_n|\boldsymbol{\eta}_n, I)] &= -\frac{1}{2}\left(\mathbf{y}_n^T\mathbf{y}_n - 2\hat{\boldsymbol{\mu}}_n^T\mathbf{y}_n + \mathrm{Tr}(\hat{\Sigma}_n) + \hat{\boldsymbol{\mu}}_n^T\hat{\boldsymbol{\mu}}_n\right) - \frac{c}{2}\log 2\pi \\ E_q[\log q(\boldsymbol{\eta}_n|\hat{\boldsymbol{\mu}}_n, \hat{\Sigma}_n)] &= -\frac{k}{2} - \frac{k}{2}\log 2\pi + \frac{1}{2}\log|\hat{\Sigma}_n^{-1}|\end{aligned}$$

The best lower bound can be obtained by maximizing each $L(\hat{\boldsymbol{\mu}}_n, \hat{\Sigma}_n, W, \Sigma)$ with respect to the variational parameters $\hat{\boldsymbol{\mu}}_n$ and $\hat{\Sigma}_n$, which gives

$$(4)\quad \hat{\boldsymbol{\mu}}_n = (\Sigma^{-1} + I)^{-1}(\Sigma^{-1}W\mathbf{x}_n + \mathbf{y}_n)$$

$$(5)\quad \hat{\Sigma}_n = (\Sigma^{-1} + I)^{-1} \ .$$

The lower bound of the log-likelihood on the whole dataset Y is given by $\sum_{n=1}^N L(\hat{\boldsymbol{\mu}}_n, \hat{\Sigma}_n, W, \Sigma)$. To obtain the estimate for model parameters, we use this lower bound function as a surrogate objective to be maximized. Given a fixed value of $(\hat{\boldsymbol{\mu}}_n^*, \hat{\Sigma}_n^*)$ from (4) and (5), the lower bound function $\sum_{n=1}^N L(\hat{\boldsymbol{\mu}}_n^*, \hat{\Sigma}_n^*, W, \Sigma)$ is a function of model parameters (W, Σ). By maximizing $\sum_{n=1}^N L(\hat{\boldsymbol{\mu}}_n^*, \hat{\Sigma}_n^*, W, \Sigma)$ with respect to W and Σ, we have

$$(6)\quad W = \left(\sum_{n=1}^N \hat{\boldsymbol{\mu}}_n\mathbf{x}_n^T\right)\left(\sum_{n=1}^N \mathbf{x}_n\mathbf{x}_n^T\right)^{-1}$$

$$(7)\quad \Sigma = \frac{1}{N}\sum_{n=1}^N \left(\hat{\Sigma}_n + (\hat{\boldsymbol{\mu}}_n - W\mathbf{x}_n)(\hat{\boldsymbol{\mu}}_n - W\mathbf{x}_n)^T\right) \ .$$

3.3. **Variational optimization.** Following the update equations in (4)-(7), we construct a variational optimization algorithm to learn the model. Starting from an initial guess of $(W^{(0)}, \Sigma^{(0)})$, the algorithm alternates between the following two steps in each iteration t:

(1) Inference-step: Given $(W^{(t-1)}, \Sigma^{(t-1)})$, for each $(\mathbf{x}_n, \mathbf{y}_n)$, find the optimal variational parameters

$$(\hat{\boldsymbol{\mu}}_n^{(t)}, \hat{\Sigma}_n^{(t)}) = \underset{(\hat{\boldsymbol{\mu}}_n, \hat{\Sigma}_n)}{\arg\max}\ L(\hat{\boldsymbol{\mu}}_n, \hat{\Sigma}_n, W^{(t-1)}, \Sigma^{(t-1)}) \ ,$$

which can be done using (4) and (5).

(2) Optimization-step: Maximizing the aggregate lower bound gives us an improved estimate of the model parameters:

$$(W^{(t)}, \Sigma^{(t)}) = \underset{(W,\Sigma)}{\arg\max} \sum_{n=1}^N L(\hat{\boldsymbol{\mu}}_n^{(t)}, \hat{\Sigma}_n^{(t)}, W, \Sigma) \ ,$$

which can be done following (6) and (7).

After t iterations, the objective function becomes $L(\hat{\boldsymbol{\mu}}_n^{(t)}, \hat{\Sigma}_n^{(t)}, W^{(t)}, \Sigma^{(t)})$. In the $(t+1)^{th}$ iteration, we have

$$\sum_{n=1}^{N} L(\hat{\boldsymbol{\mu}}_n^{(t)}, \hat{\Sigma}_n^{(t)}, W^{(t)}, \Sigma^{(t)}) \leq \sum_{n=1}^{N} L(\hat{\boldsymbol{\mu}}_n^{(t+1)}, \hat{\Sigma}_n^{(t+1)}, W^{(t)}, \Sigma^{(t)})$$
$$\leq \sum_{n=1}^{N} L(\hat{\boldsymbol{\mu}}_n^{(t+1)}, \hat{\Sigma}_n^{(t+1)}, W^{(t+1)}, \Sigma^{(t+1)}) \ .$$

The first inequality holds because $(\hat{\boldsymbol{\mu}}_n^{(t+1)}, \hat{\Sigma}_n^{(t+1)})$ maximizes $L(\hat{\boldsymbol{\mu}}_n, \hat{\Sigma}_n, W^{(t)}, \Sigma^{(t)})$ in the Inference-step. The second inequality holds because $(W^{(t+1)}, \Sigma^{(t+1)})$ maximizes $\sum_{n=1}^{N} L(\hat{\boldsymbol{\mu}}_n^{(t+1)}, \hat{\Sigma}_n^{(t+1)}, W^{(t+1)}, \Sigma^{(t+1)})$ in the Optimization-step. Therefore, the objective function is non-decreasing until convergence.

We note that the computations involved per iteration during training are scalable. Most operations involved are simple matrix multiplications or matrix-vector products. There is a matrix inversion involving a $d \times d$ matrix in (6), but since the matrix only depends on the feature vectors $\mathbf{x}_n$, the inverse can be computed offline, even before starting the iterations. The algorithm does need to invert Σ in every iteration. Since Σ is a $c \times c$ matrix where c is the number of classes, the inverse can be computed efficiently even for hundreds of classes.

3.4. **Prediction.** Assuming that Σ and W have been estimated from training data, we wish to predict the label vector $\bar{\mathbf{h}}$ for an unseen data point $\bar{\mathbf{x}}$. First note that the maximum likelihood estimate of $\bar{\boldsymbol{\eta}}$, given W and Σ is obtained by $\bar{\boldsymbol{\eta}}^* = W\bar{\mathbf{x}}$, since $\bar{\boldsymbol{\eta}} \sim N(W\bar{\mathbf{x}}, \Sigma)$. Similarly the maximum likelihood estimate for $\bar{\mathbf{y}}$ given $\bar{\boldsymbol{\eta}}$ is obtained as $\bar{\mathbf{y}}^* = \bar{\boldsymbol{\eta}}$, since $\bar{\mathbf{y}} \sim N(\bar{\boldsymbol{\eta}}, I)$. We thus formulate our prediction as follows:

$$\bar{\mathbf{y}}^* = W\bar{\mathbf{x}} \tag{8}$$

with

$$\bar{h}_i = \begin{cases} 1 & \text{if } \bar{y}_i^* > 0 \\ 0 & \text{otherwise .} \end{cases} \tag{9}$$

Effectively the prediction task in our model reduces to a matrix multiplication. For this reason our model can be seen as rather simple, and unlike most existing approaches, it can be easily used on millions of data points. Note that our model can also be interpreted as performing dimensionality reduction, whereby the matrix W incorporates information from both the observed labels and Σ.

3.5. **Relationship to Probabilistic Principal Component Analysis (PPCA).** Given high dimensional data points $\mathbf{x} \in \mathbb{R}^k$, in PPCA the objective is to obtain a lower-dimensional representation in $\mathbf{y} \in R^c$, where $c << k$. In particular the assumption is made [11]:

$$p(\mathbf{x}|\mathbf{y}, Z, \beta) = N(\mathbf{x}|Z\mathbf{y}, \beta^{-1}I) \tag{10}$$

where $Z \in \mathbb{R}^{k \times c}$, and $\beta^{-1}I$ denotes a spherical covariance matrix. PPCA proceeds by defining a prior of over $\mathbf{y}$ and integrating it out, while maximizing over Z.

While at first the assumptions that we make in BMR may appear similar, there are subtle but very important differences in our model. In our case both $\mathbf{x}$ and $\mathbf{y}$ are known. We define a mapping W from the higher dimensional space to the lower-dimensional space, and not the other way around as in PPCA. The covariance matrix Σ is not spherical in our case and is of size $c \times c$, rather than $k \times k$. Lastly in our model we introduce a latent variable $\boldsymbol{\eta}$, which connects observed $(\mathbf{x}, \mathbf{y})$ pairs.

BMR can be though of as a supervised dimensionality reduction approach where $(\mathbf{x}, \mathbf{y})$ pairs are known upfront. We learn a mapping W which best captures the observed label vectors and the underlying correlations.

3.6. **BMR for document classification.** In the generative process of Section 3.1, $\mathbf{x}_n$ could be any feature representation. In the application of document classification, instead of using the original vector of word occurrences, we opt to use the low-dimensional topic representation obtained from LDA. Most of the widely used topic models, such as Latent Dirichlet Allocation [6] and Correlated Topic Models [3], have a topic vector $\mathbf{z}_{nd}$ assigned to each of the D_n words in the document n. Given k topics, $\mathbf{z}_{nd}$ for topic i is a k-dimensional 0-1 vector with only the i^{th} dimension being 1 and others being 0. We then use $\bar{\mathbf{z}}_n = \frac{1}{D_n}\sum_{d=1}^{D_n}\mathbf{z}_{nd}$ as $\mathbf{x}_n$ in the generative process. The choice of $\bar{\mathbf{z}}_n$ is due to the following three reasons: (1) Interpretability: $\bar{\mathbf{z}}_n$ is a low-dimensional representation in the topic space. It is more interpretable than the original document representation, hence a more reasonable representation. (2) Optimality: Given $\mathbf{z}_{nd}$ for each word, the best representative is always the mean according to a wide variety of divergence functions [2]. (3) Simplicity: It is simple to take the mean of $\mathbf{z}_{nd}$ for each document. The complexity of the model would increase if we were to use other complicated transformations such as a non-linear function. (4) Efficiency: Our inference approach in any given iteration has to invert matrices of size $k \times k$. Using a lower-dimensional representation keeps the inference very efficient.

4. Empirical Evaluation

In this section we present our experimental results on both topic modeling and multi-label classification. All of our experiments were conducted on a subset of the ASRS data. In particular, 66309 reports were extracted which are labeled as anomalous events. Within these extracted reports there are 58 predefined classes. For instance "anomaly.ground-encounters.vehicle" would denote one such class name. For our topic modeling analysis, we used all 66309 reports. We refer to this data set as ASRS-66309.

Our multi-label classification results are generated by conducting 5-fold cross validation on a randomly selected subset of 10,000 reports pertaining to anomalies. The feature vectors for these 10,000 reports are obtained using LDA with number of topics assigned to 200. We refer to this data set as ASRS-10000. The size of the data set used for classification purposes is limited simply because some of the approaches that we compare against cannot easily handle much larger data sets.

4.1. **Topic Modeling Experiments.** We used LDA to extract topics from ASRS-66309. Table 1 shows some examples of obtained topics. The right column denotes a list of top-ranked words within a given topic, and the left column contains a name which is manually assigned to the topic in question. As we can see, these word lists are quite interpretable, and provide a reasonable representation for discussed topics.

Figure 3 shows the number of documents in each of the 58 classes. We can see that the classes are highly unbalanced with some classes containing more than ten thousand documents and others containing less than 50. The four largest classes are "anomaly.other-anomaly.other", "anomaly.non-adherence.published-procedure", "anomaly.non-adherence.clearance", and "anomaly.non-adherence.far", meaning that quite a few anomalies are the non adherence of prescribed procedures or clearance. The four smallest classes are "anomaly.ground-encounters.gear-up-landing", "anomaly.ground-encounters.animal", "anomaly.cabin-event.galley-fire" and "anomaly.inflight-encounter.skydivers". Judging from these names, we can see that all of them are potentially dangerous accidents, hence should rarely happen.

We investigate the relationship between 58 classes and 200 topics in ASRS-66309 data set. The number of topics was chosen upfront to be multiple times larger than the number of predefined classes. For each document, we have a posterior over all 200 topics. We assign a document to its most likely topic. Meanwhile, each document is also assigned to multiple classes. Therefore, we can count the number of the documents falling in both class s and topic i, a higher value indicates a closer relationship. Such a strategy yields a 58×200 matrix M, with $M(s, i)$ denoting the approximate relationship between class s and topic i. We visualize the matrix M in Figure 4, where a lighter color indicates a closer relationship. As we can see, there are several bright rows in the figure. The classes

maintenance on lights	light, illuminated, caution, master, lights, panel, overhead, checklist, warning, maintenance
passenger encountering turbulence	flight, passenger, attendants, seat, turbulence, seated, attendant, sign, hit, cabin
avoiding ground proximity	terrain, ground proximity warning system, warning, approach, pull, climb, received, maneuver, approximately, air traffic control
thunderstorm	heavy, rain, moderate, turbulence, area, thunderstorms, radar, due, difficult, feel
pressurization in the cabin	cabin, pressurization, descent, emergency, pressure, masks, control, oxygen, horn, passenger
avoiding collision	cessna, aircraft, evasive, collision, appeared, action, passed, avoid, directly, approximately
snow and ice	snow, conditions, braking, run way, action, poor, repeated, aircraft, ice, airport
gas leak maintenance	fuel, gauge, leak, quantity, aircraft, maintenance, tank, indicator, inoperative, problem
fire in cabin	smoke, fire, cabin, passenger, aircraft, flight, evacuate, emergency, attendant, cockpit
weather conditions on clearance	visual flight rules, instrument flight rules, airspace, airport, aircraft, traffic, flight, area, approach, conditions
taking off	tower, runway, position, control, hold aircraft, take off, clearance, final, heard
approaching destination	intersection, cross, descent, approach clear, clearance, xing, restricted, arrival, control
passenger medical emergency	passenger, medical, emergency, flight, oxygen, attendant, board, aircraft, landing, assistance
system failure	system, failure, failed, electrical, emergency, flight, aircraft, lost, problem loss
complying instructions	instructed, instructions, instruction, issued, complied, comply, immediately, received, air traffic control, acknowledged
door maintenance	door, open, closed, doors, opened, aircraft, handle, maintenance, flight, close
maintenance on tire and brake	tire, wheel, tires, aircraft, maintenance, brake, found, main, installed, change

TABLE 1. Extracted topics using LDA from ASRS database.

corresponding to these rows are "anomaly.other-anomaly.other", "anomaly.non-adherence.published-procedure", "anomaly.non-adherence.clearance", and "anomaly.non-adherence.far". These classes are the largest classes in Figure 3. Since the size of these classes is large, they have a higher chance to co-occur with the topics. This also reflects the fact that some of the larger classes include very broad types of documents. For instance anomaly.other-anomaly.other is lumping together anomalies which are not described by other predefined classes.

TABLE 2. anomaly.aircraft-equipment-problem.critical

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
engine	take off	oil	cabin	smell
landing	aircraft	engine	pressurization	smoke
emergency	knots	pressure	descent	odor
checklist	runway	repeat	emergency	cabin
failure	abort	maintenance	pressure	flight
shut	maintenance	quantity	masks	emergency
declared	engine	low	control	cockpit
shut down	aborted	shut	oxygen	electrical
single	roll	information	horn	burning
runway	gate	stated	passenger	landing

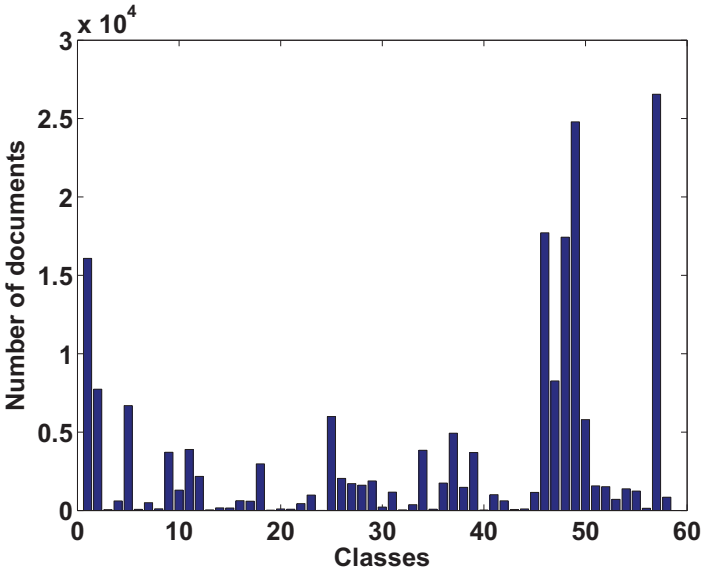


FIGURE 3. The number of documents in each of 58 classes.

TABLE 3. Top ranked topics in anomaly.excursion.taxiway

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
taxiway	aircraft	ramp	ground	snow
turn	runway	aircraft	taxiway	conditions
taxi	landing	area	control	braking
runway	touchdown	taxi	taxi	runway
taxiways	reverse	spot	runway	action
airport	normal	parking	clearance	poor
aircraft	braking	personnel	controller	repeated
area	brakes	parked	instructions	aircraft
lights	thrust	terminal	told	ice
turned	captain	turn	cleared	airport

For each class, we can rank topics according to how likely they are to occur within a given class. We examine the top ranked topics for each class. Some examples with top five topics are presented in Tables 2-6. Overall, the topic lists in each class appear reasonable. Some topics in the same class are similar to each other, and some are different but explain the class from different perspectives. For example, in Table 6, the first two topics are somewhat similar. Both of them are directly related to fire or smoke. However upon closer examination, one can see subtle differences even within these topics. The first topic appears to incorporate potential passenger attendant interactions. While the second topic includes words such as odors, smells, electrical, cockpit, indicating a potential problem in the cockpit. The third topic is related to maintenance, indicating that the system may need maintenance to avoid the fire problem. The fourth and fifth topics are related to passengers, because their misconduct, such as smoking, could be one reason for the fire.

In Table 2, for the class named critical equipment problem, we find topics on engine, maintenance, cabin pressure, and smoke. In Table 3, under taxiway excursion, we can see topics on taxiway, braking, parking, clearance, and bad weather with snow/ice. Under passenger misconduct, Table 4, we find misconduct in lavatory, cabin, security check, and also there is medical emergency and fire.

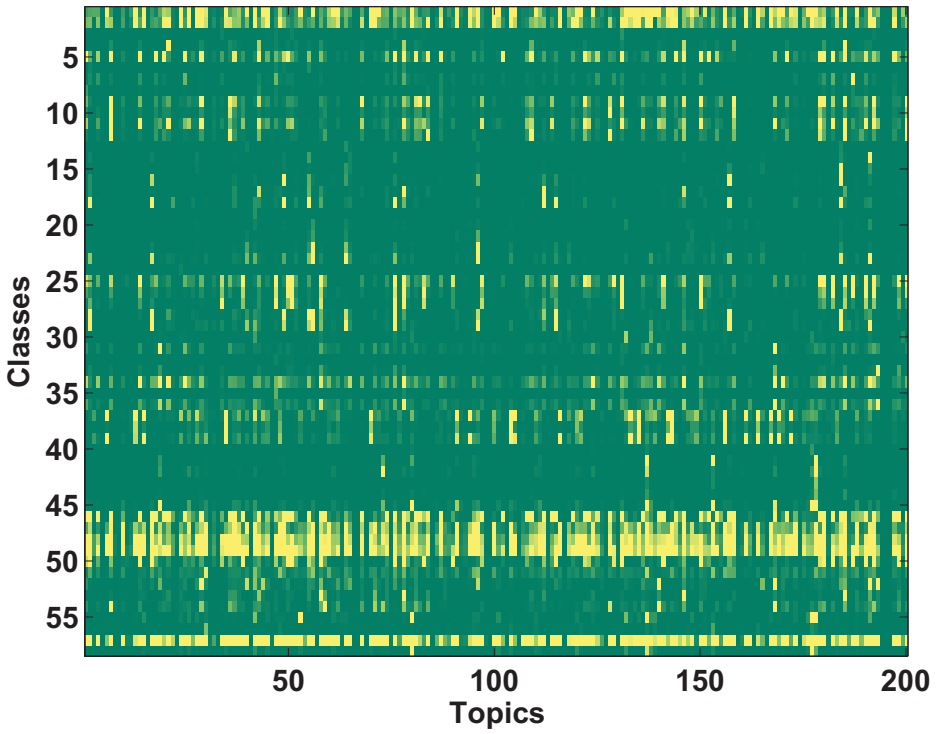


FIGURE 4. The relationship among 58 classes and 200 topics. A lighter color indicates a closer relationship.

TABLE 4. Top ranked topics in anomaly.cabin-event.passenger-misconduct

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
passenger	flight	agent	passenger	smoke
flight	attendant	passenger	medical	fire
captain	attendants	flight	emergency	cabin
seat	passenger	boarding	flight	passenger
attendant	cabin	aircraft	oxygen	aircraft
told	cockpit	security	attendant	flight
back	back	board	board	evacuate
lavatory	told	gate	aircraft	emergency
man	captain	asked	landing	attendant
purser	called	told	assistance	cockpit

In Table 5, the class of weather is associated with topics on thunderstorms, turbulence, and also landing and deviation. Overall the extracted topics do appear interpretable and reasonable.

4.2. **Multi-Label Classification Experiments.** In this section we compare the performance of our approach with existing state-of-the-art algorithms as well as baseline methods. To evaluate performance we utilize five different evaluation measures. All multi-label classification experiments are performed on the ASRS-10000 data set.

TABLE 5. Top ranked topics in anomaly.inflight-encounter.weather

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
turbulence	thunderstorms	approach	fuel	flight
moderate	deviation	runway	alternate	passenger
severe	thunderstorm	instrument landing system	air traffic control	attendants
aircraft	area	missed	emergency	seat
encountered	turn	tower	approach	turbulence
flight	due	approaches	minimum	seated
light	air traffic control	briefed	dispatch	attendant
air traffic control	avoid	landing	due	sign
repeated	emergency	final	divert	hit
ride	radar	vectors	declared	cabin

TABLE 6. Top ranked topics in anomaly.other-anomaly.smoke-or-fire

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
smoke	smell	fire	passenger	flight
fire	smoke	warning	flight	attendant
cabin	odor	engine	captain	attendants
passenger	cabin	aircraft	seat	passenger
aircraft	flight	reporter	attendant	cabin
flight	emergency	emergency	told	cockpit
evacuate	cockpit	light	back	back
emergency	electrical	checklist	lavatory	told
attendant	burning	indication	man	captain
cockpit	landing	maintenance	purser	called

4.2.1. *Algorithms.* We compare BMR with three multi-label classification algorithms. As baselines, we consider one-vs-rest SVM as a multi-label classifier, which we refer to as MLSVM. In addition we use a one-vs-rest implementation of logistic regression, which we call MLLR. We also consider two state-of-the-art approaches for multi-label learning: Multi-label K-nearest Neighbors (MLKNN) [17], a method which applies the k-nearest neighbor idea to the multi-label setting; and Instance Based Learning by Logistic Regression (IBLR) [8], where features are first transformed to incorporate label information from local neighborhoods prior to applying logistic regression.

4.2.2. *Evaluation Measures.* We evaluated performance using five different measures: one error, precision, coverage, ranking loss and hamming loss. Let $g(x, l)$ denote a real-valued function which assigns a score to label l for data point $\mathbf{x}$, such that a larger score is considered better. Also, let $f(\mathbf{x})$ denote the classifier whose output is the predicted multi-label vector. Further, let L_x denote a set of true labels associated with $\mathbf{x}$.

1) *One error* evaluates how frequently the top ranked predicted label is not among the true labels. If $[\![\cdot]\!]$ denotes the indicator function, we have:

$$OneError(g) = \frac{1}{D}\sum_{d=1}^{D} [\![\, \text{argmax}_{l\in L}\, g(\mathbf{x}_d, l) \notin L_{x_d} \,]\!] \ . \tag{11}$$

2) For true labels $l \in L_x$, *average precision* evaluates the fraction of labels in L_x that rank at least as high as l according to the scoring rule g on average. For any data point x and any label $l \in L_x$, let $\mathcal{R}(\mathbf{x}, l) = \{l' \in L_{\mathbf{x}} | rank_g(\mathbf{x}, l') \leq rank_g(\mathbf{x}, l)\}$, where the ranking is among all possible

labels. Then, average precision is:

$$AvePrec(g) = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{|L_{\mathbf{x}_d}|}\sum_{l\in L_{\mathbf{x}_d}}\frac{|\mathcal{R}(\mathbf{x}_d,l)|}{rank_g(\mathbf{x}_d,l)} \ . \tag{12}$$

3) Coverage reflects on average how far one needs to go down in the label ranking to cover all actual labels of an instance:

$$Coverage(g) = \frac{1}{D}\sum_{d=1}^{D}(\max_{l\in L_{\mathbf{x}_d}} rank_g(\mathbf{x}_d,l) - 1) \ . \tag{13}$$

4) Hamming loss evaluates the fraction of label instance pairs that were misclassified:

$$HammingLoss(f) = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{c}|f(\mathbf{x}_d)\triangle L_{\mathbf{x}_d}| \ . \tag{14}$$

where $\triangle$ denotes the symmetric difference between two sets.

5) Ranking loss reflects the average number of labels that are reversely ordered for a given instance. Let $\mathcal{T}(\mathbf{x}_d) = \{(l_1,l_2) \mid g(\mathbf{x}_d,l_1) \leq g(\mathbf{x}_d,l_2),\ (l_1,l_2) \in L_{\mathbf{x}_d}\times\bar{L}_{x_d}\}$, where $\bar{L}_{\mathbf{x}_d}$ denotes the complement of L_{x_d}. Ranking loss is defined as:

$$RankLoss(g) = \frac{1}{D}\sum_{d=1}^{D}\frac{|\mathcal{T}(\mathbf{x}_d)|}{|\bar{L}_{\mathbf{x}_d}||L_{\mathbf{x}_d}|} \ . \tag{15}$$

For both hamming loss and ranking loss, smaller values are considered better. In particular for a perfect performance $HammingLoss(h) = RankLoss(g) = 0$.

4.2.3. *Prediction Performance.* Table 7 lists the prediction results when using five fold cross validation on ASRS-10000. MLSVM and MLLR, the two one vs. rest approaches perform the worst, as expected. These results clearly illustrate that looking at hamming loss alone is actually quite misleading. For instance MLSVM has a hamming loss of 11.9%, however its one error is at 85.8%. This is especially important in ASRS, since some categories are present in only about 50 out of 66309 documents. Even for a degenerate classifier which predicts only zeros, one would obtain a low hamming loss. For this reason we have opted to evaluate our results using a range of five different evaluation measures, commonly used in multi-label classification.

Our proposed model clearly outperforms all other approaches, including MLKNN and IBLRML, the two state-of-the-art methods across all five evaluation measures. Since we have used a data set of significant size we can see that the standard deviations are quite low. It is also apparent that our improvements are indeed statistically significant. Across all evaluation measures our approached seems to be followed by IBLRML and then MLKNN. Considering the simplicity of our approach, these results are quite interesting. After all, the predictive step in our model merely involves a matrix multiplication, and yet we are outperforming very complex algorithms such as SVMs or even state-of-the-art multi-label learning methods such as MLKNN and IBLRML.

For the top three algorithms, BMR, MLKNN and IBLRML, we also examined what happens when a smaller fraction of the data set is labeled. We omitted the one vs. rest approaches to prevent clutter, and also since we already established that their performance is substantially inferior. We ran 5-fold cross validation on the ASRS-10000 data set, while gradually increasing the set of labeled points from 3000 to 4000. Since the number of classes is rather large we did not consider smaller sets. The results can be seen in Figure 5. The first thing that we can note is that the performance of IBLRML appears to be worse than that of MLKNN when the set of labeled points is smaller. However that is not the case when full 5-fold cross validation is performed (see Table 7). It appears that IBLRML requires a larger training set to achieve a good performance. Across all evaluation measures our proposed method, BMR, consistently outperforms both MLKNN and IBLRML. This

TABLE 7. Five-fold cross validation on the ASRS-10000 data set. MBR clearly outperforms all the other methods.

	BMR	MLKNN	IBLRML	MLLR	MLSVM
OneError	$\mathbf{38.5 \pm 0.8}$	44.1 ± 0.7	44.3 ± 1.4	50.7 ± 1.6	85.8 ± 18.1
AvePrec	$\mathbf{64.0 \pm 0.5}$	59.9 ± 0.5	60.3 ± 0.6	57.0 ± 0.9	33.6 ± 8.2
Coverage	$\mathbf{8.17 \pm 0.14}$	9.20 ± 0.12	8.39 ± 0.29	9.63 ± 0.51	13.81 ± 0.87
HammingLoss	$\mathbf{4.4 \pm 0.0}$	4.6 ± 0.1	4.7 ± 0.0	5.5 ± 0.1	11.9 ± 1.1
RankLoss	$\mathbf{5.7 \pm 0.2}$	6.9 ± 0.1	6.7 ± 0.3	7.9 ± 0.6	12.9 ± 1.7

seems to indicate that our approach is robust with respect to the ASRS data, even when the size of the training set is reduced.

4.2.4. *Scalability.* To contrast the computational cost involved in utilizing the MLKNN, IBLRML and MBR we conducted an experiment in which we tested how long it takes to predict on data sets between 1000 and 14000 data points. The MLKNN approach requires K-nearest neighbor computations, as such it is the most expensive. IBLRML on the other hand constructs 58 separate logistic regression classifiers and has to utilize each one of them. Figure 6 illustrates that our proposed approach is clearly the most efficient.

5. Conclusion

In this paper, we have analyzed the ASRS data from two aspects. First, we applied Latent Dirichlet Allocation to automatically extract the topics from reports in ASRS database. We have established that the topics returned by LDA are indeed quite interpretable when it comes to the ASRS data, and that they can be used to reason about potential problems that are being discussed. In particular we could see that extracted topics within each predefined category are indeed similar as one would expect. We have also successfully utilized LDA to obtain a lower-dimensional feature representation for our subsequent classification task.

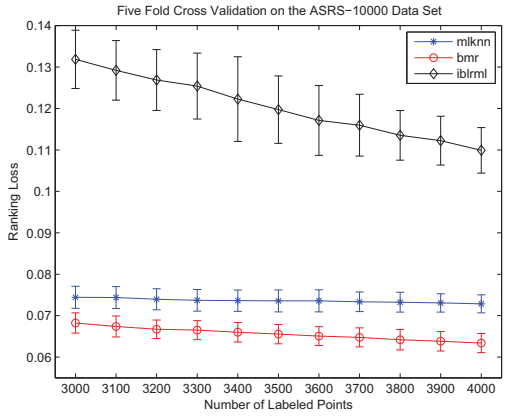
The second aspect that we have addressed involves multi-label classification. We have proposed Bayesian Multivariate Regression (BMR), a novel multi-label classification algorithm, which explicitly models the correlation structure among labels. As illustrated by our empirical evaluation our model is very effective and competitive with the state of the art across several evaluation measures, at the same time it is simple enough that it could potentially be applied to millions of data points. The scalability is possible since the learning step only involves matrix multiplications and inverting small matrices and the prediction step involves only a matrix multiplication. While we have explored this algorithm in the domain of ASRS data, its applicability extends to any domain where correlated multi-label prediction problems occur.

For future work, we intend to create a joint model which combines BMR and topic modeling. As illustrated in [14] creating a joint model may lead to even better performance. It will also be interesting to further explore BMR from the perspective of supervised dimensionality reduction.

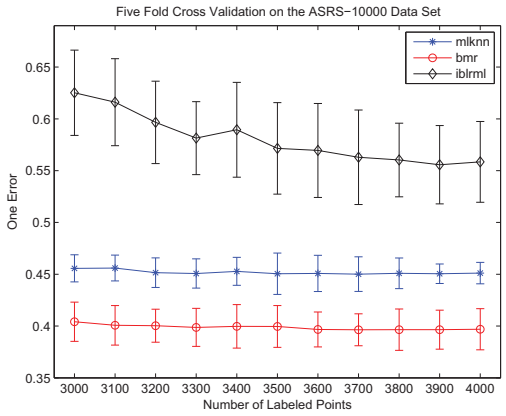
Acknowledgements. This research was supported by NASA grant NNX08AC36A, NSF grants IIS-0916750, IIS-0812183, and NSF CAREER grant IIS-0953274.

References

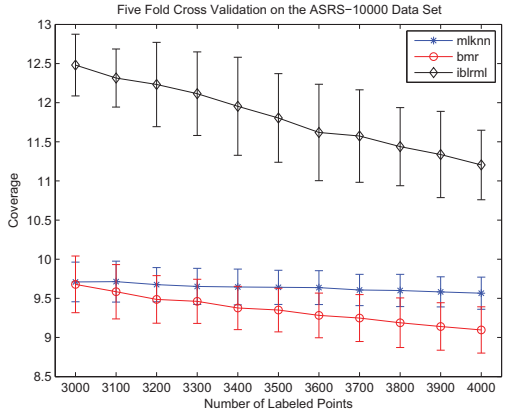
[1] Aviation Safety Reporting System. http://akama.arc.nasa.gov/ASRSDBOnline/QueryWizard_Filter.aspx.
[2] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with Bregman divergences. 6:1705–1749, 2005.
[3] D. Blei and J. Lafferty. Correlated topic models. *NIPS*, 2006.
[4] D. Blei and J. Lafferty. Dynamic topic models. In *ICML*, 2006.
[5] D. Blei and J. McAuliffe. Supervised topic models. In *NIPS*, 2007.
[6] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.



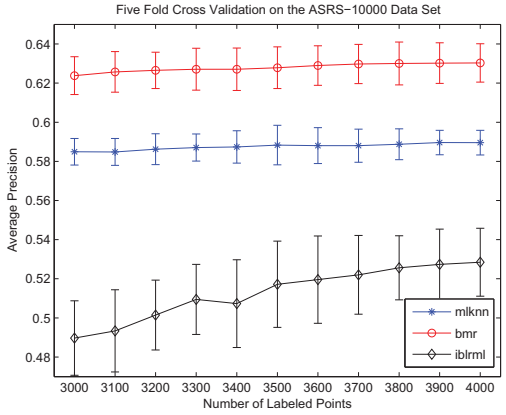
(a) Ranking Loss



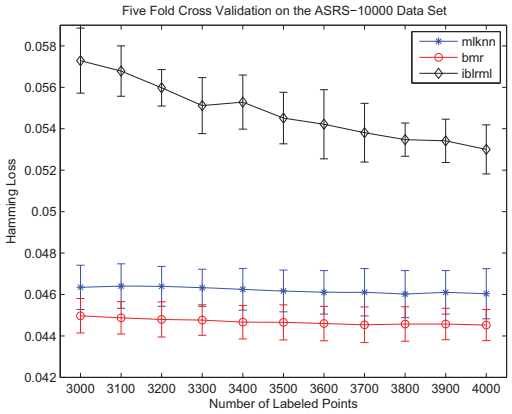
(b) One Error



(c) Coverage



(d) Average Precision



(e) Hamming Loss

FIGURE 5. Five fold cross validation on ASRS-10000 data set. To avoid clutter we only include the top three algorithms. These plots indicate what happens when a smaller fraction of the data set is labeled. Even in this setting BMR consistently outperforms both MLKNN and IBLRML.

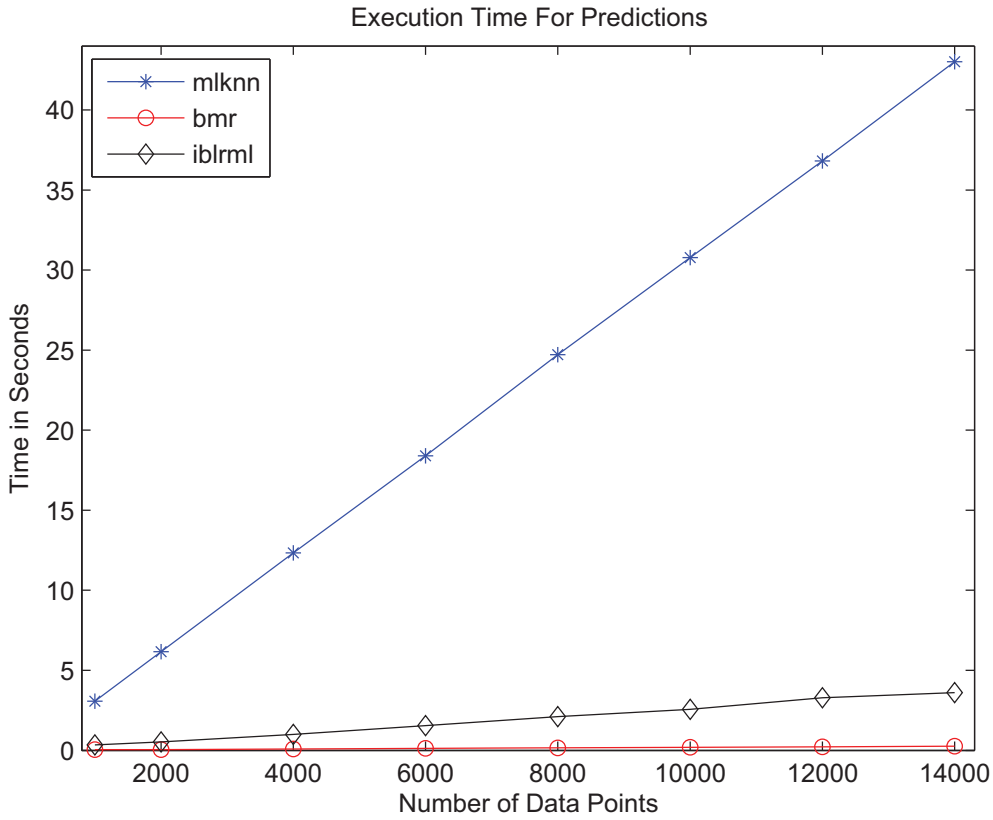


FIGURE 6. Computational time to make predictions as more and more points are considered. MBR outperforms MLKNN and IBLRML.

[7] G. Chen, Y. Song, F. Wang, and C. Zhang. Semi-supervised multi-label learning by solving a sylvester equation. In *SDM*, 2008.
[8] W. Cheng and E. Hüllermeier. Combining instance-based learning and logistic regression for multilabel classification. *Mach. Learn.*, 76(2-3):211–225, 2009.
[9] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *ACM SIGIR*, 2005.
[10] S. Ji and J. Ye. Linear dimensionality reduction for multi-label classification. In *IJCAI*, 2009.
[11] N. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *NIPS*, 2003.
[12] A. McCallum. Multi-label text classification with a mixture model trained by EM. In *AAAI Workshop on Text Learning*, 1999.
[13] P. Rai and H. Daume. Multi-label prediction via sparse infinite CCA. In *NIPS*, 2009.
[14] H. Shan and A. Banerjee. Discriminative mixed-membership models. In *ICDM*, 2009.
[15] Y. Song, L. Zhang, and L. Giles. A sparse gaussian processes classification framework for fast tag suggestions. In *CIKM*, 2008.
[16] L. Tang, J. Chen, and J. Ye. On multiple kernel learning with multiple labels. In *IJCAI*, 2009.
[17] M. Zhang and Z. Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.
[18] Y. Zhang and Z. Zhou. Multi-label dimensionality reduction via dependence maximization. In *AAAI*, 2008.