Multi-labelled Classification Using Maximum Entropy Method

Shenghuo Zhu Xiang Ji Wei Xu Yihong Gong
{zsh,xji,xw,ygong}@sv.nec-labs.com
NEC Laboratories America, Inc.
10080 North Wolfe Road SW3-350
Cupertino, CA 95014

ABSTRACT

Many classification problems require classifiers to assign each single document into more than one category, which is called *multi-labelled classification*. The categories in such problems usually are neither conditionally independent from each other nor mutually exclusive, therefore it is not trivial to directly employ state-of-the-art classification algorithms without losing information of relation among categories. In this paper, we explore correlations among categories with maximum entropy method and derive a classification algorithm for multi-labelled documents. Our experiments show that this method significantly outperforms the combination of single label approach.

Categories and Subject Descriptors: H.3.3 [Information Systems]: Information Search and Retrieval

General Terms: Algorithms, Experimentation

Keywords: multi-labelled classification, maximum entropy method

1. INTRODUCTION

Data classification is the task of assigning each of the given data to a set of predefined categories. In general, all classification problems can be categorized as either single-labelled, or multi-labelled problems. Single-labelled data classification assumes that the predefined data categories are mutually exclusive and each data point can belong to exactly one category. Binary classification is the simplest case of the single-labelled problem where each data point is assigned to one of two predefined categories. To date, many classification methods, such as Naive Bayes, SVM, and Logistic Regression, have been developed to address the single-labelled classification problem. On the other hand, with multi-labelled classification, the data categories may not be either mutually exclusive or conditionally independent, and each data point can belong to multiple categories simultaneously. Multi-labelled classification problems are very common in the areas of document analysis and information retrieval. For example, a newspaper article about the presidential election may talk about a wide range of topics such as politics, economy, and foreign relations; an email discussing the ongoing business work may also include topics about the past vacation the sender had experienced with his friends; etc. For document retrieval, a user may want to retrieve the news simultaneously belonging to multiple categories, which requires classifiers to correctly assign documents to all categories.

Despite the value and the significance of the problem, research on multi-labelled classification has received much less attention compared to its single-labelled counterpart. Currently the most common solution to the multi-labelled classification problem is to decompose the problem into multiple, independent binary classification problems, and determine the final labels for each data point by aggregating the classification results from all the binary classifiers. More precisely, for a given m predefined categories, m binary classifiers are independently created, one for each category, and are used to determine if a given data point belongs to the corresponding category or not. The final category label for the data point is determined by combining the category labels generated by these m binary classifiers. The advantage of this approach is that a multi-labelled classifier can be readily built using many start-of-the-art binary classifiers off the shelf, such as SVM. However, when there exist strong correlations among categories, data classification performance may deteriorate because this approach employs a set of independent binary classifiers to conduct data classifications, and mutual correlations among different categories are completely ignored. More specifically, given the input variables, the optimal estimate should be the labels with the largest joint probability, instead of the combination of labels with largest individual probabilities of categories. Later we illustrate the difference in Section 3.2.

To take the dependencies among data categories into account, a straight-forward approach is to transform the multi-labelled classification problem into a single-labelled problem by treating each possible combination of categories as a new class. In other words, a multi-labelled classification problem with ten predefined classes would be transformed to a single-labelled classification problem with 1024 classes each of which corresponds to a possible combination of the original data classes. However, this approach faces the problem of data sparseness because there could be very few data points in many combinations of the data classes.

In this paper, we propose a multi-labelled data classification method by explicitly modelling the mutual correlations among data categories using the maximum entropy principle. Our method accomplishes the multi-labelled data classification task by constructing a conditional probability model $\Pr(\mathbf{y}|\mathbf{x})$ from the training data set, where $\mathbf{x}$ is the feature vector of the input data point, and $\mathbf{y}$ is the class membership vector in which each element y_i indicates whether $\mathbf{x}$ belongs to the i'th class or not. In contrast to traditional approaches where $\Pr(\mathbf{y}|\mathbf{x})$ is usually determined by the class pri-

SIGIR'05, August 15–19, 2005, Salvador, Brazil.

ors and feature vectors of the input data, we construct $\Pr(\mathbf{y}|\mathbf{x})$ by including an additional term — the dependencies among the data classes. We employ the Maximum Entropy (ME) method to estimate the parameters during the model construction process. To reflect the estimation errors between the empirical and the real distributions, we introduce the regularization parameters, which serve to avoid the over-fitting problem for the model construction. This measure is in analogy to the penalized logistic regression, except for the items serving for the correlation among category labels. Our experimental evaluations show that the proposed multi-labelled classification method reveals statistically significant performance improvements compared to traditional approaches.

The remainder of the paper is organized as follows. The related work is discussed in Section 2. In Section 3, we describe the model of multi-label maximum entropy. Then, we present the experiments and results in Section 4. Finally Section 5 concludes the paper.

2. RELATED WORK

There is limited work on the problem of multi-labelled classification. In the literature, many research studies take ranking-based approaches which assign a real-valued score to each document-category pair, and classify each document by choosing all the categories with the scores above the given threshold. Schapire and Singer [16] proposed BoosTexter which essentially is an enhancement to AdaBoost to build the ranks for all document-category pairs by using the boosting techniques. Elisseeff and Weston [8] developed a method using a kernel SVM as the ranking function for document-category pairs. Crammer and Singer [6] proposed a family of one-against-rest online ranking algorithms that create a weight vector for each category, and compute the ranking between a document and a category using the inner product of the document's feature vector and the category's weight vector.

Although ranking-based approaches provide a unique way to handle the multi-labelled classification problem, they generally do not explicitly model the correlations among data categories. Another problem such methods are facing with is that it is difficult to determine into how many categories a particular data should be classified, and thresholds are usually selected heuristically.

Another common approach to the multi-labelled classification problem is the modeling of classification using generative probabilistic models. McCallum [14] described a method based on generative model which assumes that each multi-labelled document is generated by a mixture of single-labelled document models. The method resorts to naïve Bayes model for each category model by assuming the independence between words given category. The method employs the expectation-maximization (EM) to estimate the model parameters and the mixture parameters. Ueda and Saito [17] also proposed a probabilistic generative model that uses a different mixture approach. The advantage of these methods is that they explicitly model the category correlations, and require no threshold for determining the category label for each data point. However, because these methods usually assume words independence and mixture of category features within their probabilistic models, data classification accuracies could be limited because these assumptions usually do not reflect the real-world data configurations.

A closely related approach in the literature is the one proposed by Godbole and Sarawagi [10] that stacks two levels of SVM's with heterogeneous features. Each lower level SVM is a single-labelled, one-against-rest classifier with the original text features as the input. Combining the original text features, the outputs of the lower level SVM's are used as the input of the higher level SVM's which determine the final category label for each document. In Section 4, we implement a variation of this method, and compare it with our proposed multi-labelled data classification method for performance evaluations.

In addition, there is some other work closely related to the multi-labelled classification problem. Clare and King [4] developed a method, which uses a modified entropy measure to extend the algorithm C4.5 to allow nodes containing multiple labels. The method also uses resampling strategies to deal with classes with small numbers of examples. Similarly, Comite et al. [5] extended the alternating decision trees (ADTrees) algorithm for multi-labelled problems. Each node of their multi-label ADTrees is associated with a set of real values, one for each label. Har-Peled et al. [11] described a constraint classification framework. Under the framework, classification problems are translated into a binary classification in a higher dimensional space with certain constraints. The paper also presented a meta-learning algorithm that learns via a single linear classifier consistent with the constraints. However, the correlations among labels were not explicitly discussed in the paper. Cai and Hofmann [2] proposed a hierarchical approach for multi-labelled/multi-class classification problem, where the predefined taxonomy is used to redefine the loss functions. Gao et al. [9] extended their binary maximal figure-of-merit learning algorithm to multi-labelled classification problem. The method optimizes the performance against the approximated evaluation criteria, but the discriminant function for classification is still based on individual categories.

3. THE MULTI-LABELLED MAXIMUM ENTROPY METHOD

Here we briefly review the method of single-labelled data classification using the maximum entropy model. When we consider the relaxation of the constraints, it is equivalent to the penalized logistic regression method. Then, we give an example to demonstrate why the approach of combining single-labelled classifiers does not work well for the multi-labelled classification problem, which implies the importance of modeling the dependency among data categories. Finally, we propose the multi-labelled maximum entropy approach to model the dependency among category labels.

Let $\mathbf{x} = (x_1, \cdots, x_d)^\top$ denote the random variable representing feature vectors of the input data; let $\mathbf{y}$ denote the category label vector of a particular data point (we describe the details of $\mathbf{y}$ in different situations). Statistical approaches accomplish the data classification task by estimating the conditional probability $\Pr(\mathbf{y}|\mathbf{x})$ from the training data, and determine the category label $\hat{\mathbf{y}}$ of a given data point with the feature vector $\mathbf{x}$ using the following equation:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{Y}} \Pr(\mathbf{y}|\mathbf{x}). \quad (1)$$

where $\mathcal{Y}$ represents the label space of the entire data set.

3.1 Maximum entropy method for single-labelled classification

In this section, we briefly introduce the single-labelled data classification using the maximum entropy model. Detailed descriptions of the method can be found in [15]. For simplicity, we only describe the binary classification case. Therefore, the label space $\mathcal{Y} = \mathfrak{B}$, where $\mathfrak{B}$ is a binary space, containing 0 and 1[1]. Since $\mathbf{y}$ has only one dimension, we denote it as y.

The principle of the *maximum entropy model* (MEM) [12] is simple: model all that is known and assume nothing about what is

[1]In literature, some use -1 and 1. Two representations are equivalent except for resulting different parameter values.

unknown. In other words, given a collection of facts, the MEM chooses a model which is consistent with all the facts, but otherwise is as *uniform* as possible. In real implementations, facts are usually represented as a set of constraints, and the optimal model is acquired by maximizing the model's entropy under the given constraints.

Let $\tilde{P}(\mathbf{x}, y)$, $Q(\mathbf{x}, y)$ denote the empirical and the model distributions, respectively. Traditional MEM-based data classification methods typically use the following constraints for model selection:

$$\begin{aligned} \langle y \rangle_Q &= \langle y \rangle_{\tilde{P}}, \\ \langle y x_l \rangle_Q &= \langle y x_l \rangle_{\tilde{P}}, \quad \forall 1 \le l \le d, \end{aligned} \tag{2}$$

where $\langle \cdot \rangle_P$ denotes the expectation with respect to distribution P; x_l represents an element of the feature vector $\mathbf{x}$. The above two constraints serve to force the model under construction to comply with the two statistical properties of the training data set: the prior probability of each category, and the correlations among the categories and features of the given data.

For the problem of data classification, the model to be estimated is the conditional probability $Q(y|\mathbf{x})$ (denoted as a function of y and $\mathbf{x}$, $q(y|\mathbf{x})$, from now on) and the MEM obtains the optimal $q(y|\mathbf{x})$ by maximizing the following entropy subject to the constraints Eq. (2) and $\sum_y q(y|\mathbf{x}) = 1$. We have

$$\hat{q} = \arg\max_q \mathcal{H}(\mathbf{x}, y|Q) = \arg\min_q \langle \log q(y|\mathbf{x}) \rangle_Q, \tag{3}$$

where $\mathcal{H}(\mathbf{x}, y|Q)$ is the entropy of $\mathbf{x}$ and y given distribution Q with parameter q. By expanding $\mathcal{H}(\mathbf{x}, y|Q)$ and ignoring the constants irrelevant to $q(q|\mathbf{x})$, we have the second part of Eq. (3).

The minimization of Eq. (3) is a typical constrained optmization problem that can be solved using Lagrange Multiplier algorithms. The Lagrangian of Eq. (3) is:

$$\begin{aligned} \mathcal{L}(q(y|\mathbf{x}), b, \mathbf{w}, \zeta(\mathbf{x})) &= \langle \log q(y|\mathbf{x}) \rangle_Q \\ &+ b(\langle y \rangle_{\tilde{P}} - \langle y \rangle_Q) + \sum_l w_l (\langle y x_l \rangle_{\tilde{P}} - \langle y x_l \rangle_Q) \\ &+ \sum_{\mathbf{x}} \zeta(\mathbf{x})(1 - \sum_y q(y|\mathbf{x})), \end{aligned} \tag{4}$$

where b, $\mathbf{w} = (w_1, \cdots, w_d)^\top$ and $\zeta(\mathbf{x})$ are the Lagrangian multipliers. Omitting the mathematical derivations (refer to [12, 7] for derivation details), the optimal model $\hat{q}$ takes the form of

$$\hat{q}(y|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp(y(b + \mathbf{w}^\top \mathbf{x})), \tag{5}$$

where $Z(\mathbf{x}) = \sum_y \exp(y(b + \mathbf{w}^\top \mathbf{x}))$ is the partition function.

The constraints in Eq. (2) assume the model distribution equals the empirical distribution. However, for a limited number of training data, there exist estimation errors. Without considering such errors, the solution may lead to generation errors. To have a robust estimation, Chen and Rosenfeld [3] proposed to introduce maximum a posteriori probability (MAP) model under the Gaussian prior into the constraints. Assuming η and ϕ_l are the estimate errors which follow Gaussian distributions with zero means and variances of σ_η^2/n and σ_ϕ^2/n (n is the number of documents), respectively, we rewrite Eq. (2) as Eq. (6).

$$\begin{aligned} &\langle y \rangle_Q = \langle y \rangle_{\tilde{P}} + \eta, \\ &\langle y x_l \rangle_Q = \langle y x_l \rangle_{\tilde{P}} + \phi_l, \quad \forall 1 \le l \le d, \\ &\frac{\eta^2}{2\sigma_\eta^2/n} + \sum_l \frac{\phi_l^2}{2\sigma_\phi^2/n} \le C, \end{aligned} \tag{6}$$

where C is a parameter that can be used to set the tolerance of the estimation errors.

With the renewed constraints, the Lagrangian becomes:

$$\begin{aligned} \mathcal{L}(q(y|\mathbf{x}), \eta, \phi, b, \mathbf{w}, \gamma, \zeta(\mathbf{x})) &= \langle \log q(y|\mathbf{x}) \rangle_Q \\ &+ b(\langle y \rangle_{\tilde{P}} + \eta - \langle y \rangle_Q) + \sum_l w_l (\langle y x_l \rangle_{\tilde{P}} + \phi_l - \langle y x_l \rangle_Q) \\ &+ \gamma(\frac{\eta^2}{2\sigma_\eta^2/n} + \sum_l \frac{\phi_l^2}{2\sigma_\phi^2/n} - C) \\ &+ \sum_{\mathbf{x}} \zeta(\mathbf{x})(1 - \sum_y q(y|\mathbf{x})), \end{aligned} \tag{7}$$

where b, $\mathbf{w} = (w_1, \cdots, w_d)^\top$, γ (≥ 0), and $\zeta(\mathbf{x})$ are the Lagrangian multipliers.

By solving Eq. (7) and ignoring constants, we have

$$\begin{aligned} \mathcal{L}(b, \mathbf{w}) = \Big\langle -y(b + \mathbf{w}^\top \mathbf{x}) + \log Z(\mathbf{x}) \Big\rangle_{\tilde{P}} \\ + \frac{\lambda_b}{2n} b^2 + \frac{\lambda_\mathbf{w}}{2n} ||\mathbf{w}||_2^2, \end{aligned} \tag{8}$$

where $\lambda_b = \sigma_\eta^2/\gamma$ and $\lambda_\mathbf{w} = \sigma_\phi^2/\gamma$, $||\cdot||_2$ denotes the 2-norm. Actually, λ_b and $\lambda_\mathbf{w}$ serve as regularization coefficients for the bias term and the feature terms, respectively. In many applications, the bias term is not regularized, which means to set λ_b to zero. In [20], by adding a constant feature, the bias term is treated the same as the feature terms, which is equivalent to $\lambda_b = \lambda_\mathbf{w}$. Actually, when there are a large number of training data, the difference between these two settings is very small. In our experiments, we set $\lambda_b = 0$.

Eq. (8) is actually penalized logistic regression (cf. [21]). The classification task is to find the optimal parameters $\hat{b}$ and $\hat{\mathbf{w}}$ to minimize $\mathcal{L}(b, \mathbf{w})$ in Eq. (8). Plugging the optimal parameters, $\hat{b}$ and $\hat{\mathbf{w}}$, into Eq. (5), we have optimal conditional distribution $\hat{q}(y|\mathbf{x})$, which is used to classify a given document with feature vector $\mathbf{x}$.

3.2 Why not combine single labels

For the multi-labelled classification problem, let $\mathbf{y} = (y_1, \cdots, y_m)^\top \in \mathcal{Y} \subset \mathfrak{B}^m$ be the label vector of a data point, where m is the total number of categories, and each dimension y_i of $\mathbf{y}$ indicates the membership of the data point in category i. By assuming the independence among the categories, the approach of combining single-labelled classifiers for multi-labelled data classification can be expressed as follows:

$$\begin{aligned} \hat{\mathbf{y}} &= \arg\max_{\mathbf{y} \in \mathcal{Y}} \prod_{i=1}^m \Pr(y_i|\mathbf{x}) \\ &= \Big(\cdots, \arg\max_{y_i \in \mathfrak{B}} \Pr(y_i|\mathbf{x}), \cdots \Big). \end{aligned} \tag{9}$$

The following example shows why combining single-labelled classifiers does not always produce correct results for the multi-labelled classification problem when the categories are not independent. Assume that the joint distribution $\Pr(y_1, y_2|\mathbf{x})$ for some data point $\mathbf{x}$ is shown in Table 1. Further assume that we trained

$\Pr(y_1, y_2\|\mathbf{x})$	$y_1 = 0$	$y_1 = 1$	$\Pr(y_2\|\mathbf{x})$
$y_2 = 0$	0	**0.4**	0.4
$y_2 = 1$	0.3	0.3	**0.6**
$\Pr(y_1\|\mathbf{x})$	0.3	**0.7**	

Table 1: An example of joint distribution of two labels.

two single-labelled classifiers independently, which yields the conditional probabilities $\Pr(y_1|\mathbf{x})$ and $\Pr(y_2|\mathbf{x})$ shown in the same table. Because $\Pr(y_1 = 0|\mathbf{x}) = 0.3$ is less that $\Pr(y_1 = 1|\mathbf{x}) = 0.7$, data x is assigned to the first category $y_1 = 1$. Similarly, data x is assigned to the second category $y_2 = 1$ as well. However, according to Table 1, $\Pr(y_1 = 1, y_2 = 1|\mathbf{x}) = 0.3$, which is less than $\Pr(y_1 = 1, y_2 = 0|\mathbf{x}) = 0.4$. This means that the correct category labels for data $\mathbf{x}$ is $y_1 = 1, y_2 = 0$, and the result generated by combining the two single-labelled classifiers is not correct!

Clearly, the approach of combining single-labelled classifiers without considering the dependence among category labels has its limitation on the multi-labelled classification problem. Therefore, we develop a multi-labelled data classifier using the maximum entropy model in the following section.

3.3 Multi-labelled maximum entropy model

For the multi-labelled classification problem, we can extend the constraints in Eq. (6) to

$$\begin{aligned} \langle y_i \rangle_Q &= \langle y_i \rangle_{\tilde{P}} + \eta_i, \quad \forall 1 \le i \le m, \\ \langle y_i x_l \rangle_Q &= \langle y_i x_l \rangle_{\tilde{P}} + \phi_{il}, \quad \forall 1 \le i \le m, 1 \le l \le d, \end{aligned} \tag{10}$$

where η's and ϕ's are estimate errors.

As the previous example shows, correlations among categories are important to the multi-labelled classification problem. To capture such information, we add a new type of constraints to the maximum entropy model to require the model to comply with the second order statistical property $y_i y_j$ of the training data.

$$\langle y_i y_j \rangle_Q = \langle y_i y_j \rangle_{\tilde{P}} + \theta_{ij}, \quad \forall 1 \le i < j \le m, \tag{11}$$

where θ's are estimate errors.

Although it is possible to use other higher order statistics to model the category dependencies, the cost of employing such statistics may surpass the benefits they bring about. The higher order the statistics, the more parameters the model needs to estimate. With limited training data, models involving higher order statistics can hardly capture true distributions of the underlying data, and are likely to end up with little difference or even deteriorated performances compared to models using lower order statistics.

Again, the problem in our hands is to obtain the optimal $q(\mathbf{y}|\mathbf{x})$ that maximizes the entropy in Eq. (3) subject to the constraints in Eq. (10), (11) and $\sum_{\mathbf{y}} q(\mathbf{y}|\mathbf{x}) = 1$. Similar to Eq. (5) (See Appendix for the derivation details), we have

$$\hat{q}(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp(\mathbf{y}^\top (\mathbf{b} + R\mathbf{y} + W\mathbf{x})), \tag{12}$$

where $Z(\mathbf{x}) = \sum_{\mathbf{y}} \exp(\mathbf{y}^\top (\mathbf{b} + R\mathbf{y} + W\mathbf{x}))$ is the partition function; $\mathbf{b} = (b_1, \cdots, b_m)^\top$, W (an $m \times d$ matrix), and R (an $m \times m$ strict upper triangle matrix) are Lagrangian multipliers that need to be determined. By simplifying the Lagrangian and ignoring constants, we have

$$\begin{aligned} \mathcal{L}(\mathbf{b}, R, W) = &\left\langle -\mathbf{y}^\top (\mathbf{b} + R\mathbf{y} + W\mathbf{x}) + \log Z(\mathbf{x}) \right\rangle_{\tilde{P}} \\ &+ \frac{\lambda_{\mathbf{b}}}{2n} ||\mathbf{b}||_2^2 + \frac{\lambda_R}{2n} ||R||_F^2 + \frac{\lambda_W}{2n} ||W||_F^2, \end{aligned} \tag{13}$$

where $||\cdot||_F$ denotes Frobenius norm, $\lambda_{\mathbf{b}} = \sigma_\eta^2/\gamma$, $\lambda_R = \sigma_\theta^2/\gamma$ and $\lambda_W = \sigma_\phi^2/\gamma$. Similar to Eq.(8), $\lambda_{\mathbf{b}}$, λ_R and λ_W act as regularization coefficients, and there values are to be specified by the user.

Here, the task of finding the optimal $\hat{q}(\mathbf{y}|\mathbf{x})$ becomes the problem of finding the optimal $\mathbf{b}$, W, and R that minimizes the Lagrangian:

$$\hat{\mathbf{b}}, \hat{R}, \hat{W} = \arg\min_{\mathbf{b}, R, W} \mathcal{L}(\mathbf{b}, R, W). \tag{14}$$

Eq. (14) can be solved using gradient descent approaches. The derivatives of $\mathcal{L}$ with respect to its parameters are

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial b_i} &= \langle y_i \rangle_Q - \langle y_i \rangle_{\tilde{P}} + \frac{\lambda_{\mathbf{b}}}{n} b_i, \\ \frac{\partial \mathcal{L}}{\partial R_{ij}} &= \langle y_i y_j \rangle_Q - \langle y_i y_j \rangle_{\tilde{P}} + \frac{\lambda_R}{n} R_{ij}, \\ \frac{\partial \mathcal{L}}{\partial W_{il}} &= \langle y_i x_l \rangle_Q - \langle y_i x_l \rangle_{\tilde{P}} + \frac{\lambda_W}{n} W_{il}. \end{aligned}$$

There are many gradient descent methods off the shelf. In [13], Malouf compared several algorithms for maximum entropy parameter estimation and suggests that the limited memory variable metric (LMVM) [1] method is the fastest solver for document classification problems. Therefore, in our implementation we use LMVM to estimate the parameters.

Once we have $\hat{\mathbf{b}}, \hat{R}, \hat{W}$, classifying a document with feature vector $\mathbf{x}$ is equivalent to

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{Y}} \mathbf{y}^\top (\hat{\mathbf{b}} + \hat{R}\mathbf{y} + \hat{W}\mathbf{x}). \tag{15}$$

To label a data point, we can enumerate all possible label sets in $\mathcal{Y}$ to find the most probable one using Eq.(15).

4. EXPERIMENTS

To show the benefit of using multi-labelled maximum entropy method, we evaluate the method against other methods on two real data sets.

4.1 Data description

The first data set is the Reuters-21578 document corpus that contains 21578 documents collected from the Reuters newswire in 1987. It is a standard text categorization benchmark test set that consists of 135 document categories. In our experiments, we used the ten (10) largest categories for performance evaluations. Table 2 shows the statistics of document labels in our training set. It is observed from the table that only 6.5% of the documents in the training set possess multiple labels (i.e., belong to multiple categories).

No. of labels	No. of documents	percentage
0	3113	32.4%
1	5870	61.1%
2	542	5.6%
3	73	0.8%
4	5	0.1%
total	9603	100%

Table 2: Numbers of multi-labelled document in the training set of Reuters-21578 data set.

To prepare the features for documents, we follow the widely used bag-of-word approach. The features used in our experiments are words that appear more than once in the corpus. All the documents are processed with the following steps: removing SGML tags, downcasing, removing words on the SMART stoplist, stemming. The above pre-processing has resulted in a total of 11084 words as the final features. We employed the TFIDF weighting scheme and the normalization in creating the feature vector for each document. We used the modified Apte ("ModApte") split to create

the training and the testing sets that consist of 9603 and 3299 documents, respectively.

The second data set is an email corpus collected by us from six public domain mailing lists[2]. Our original purpose for creating such an email corpus is to monitor the R&D activities of a project group and discover the contributions of each employee through mining and analysis of emails among the group members. To serve these purposes, we have defined the following nine categories for email classification: (1) Topic Raising (RAISE), (2) Question Asking (ASK), (3) Work Report (REP), (4) Information Announcement(INFO), (5) Delegation (DEL), (6) Solution Proposal (SP), (7) Positive Comments (POSCOM), (8) Negative Comments (NEGCOM), and (9) Others (OTHERS). Our pre-processing on the email corpus includes removal of irrelevant information and extraction of implicit features. We remove the following items from the body of each email: attachments (pictures, executable codes), marker characters, quoted materials, email header, signature, time information, reply information, debug message, compiling message, source codes, etc. The extracted implicit features include: reply relation, reply indicator, hyper-links, ftp sites, itemization symbols, "forwarded" mark in email title, type of attached data, etc. In our email corpus, a large percentage of emails are assigned with multi-labels. For example, 57.1% (474/830) of the emails in class RAISE also belong to class ASK; 41.2% (474/1150) of the emails in class ASK also belong to class RAISE. In our experiments, we use the first eight (8) categories, and treat emails in the OTHERS category as having no labels. We found that 34.6% of documents in the training set have more than one label (see Table 3). The percentage of multi-labelled documents in our email corpus is significantly higher than that of the Reuters-21578 corpus. The procedures used for creating the feature vector of each email are similar to those for the Reuters-21578 corpus, which results in a total of 3947 words as the final feature set.

No. of labels	No. of documents	percentage
0	91	2.4%
1	2363	63.0%
2	1104	29.4%
3	186	5.0%
4	9	0.2%
total	3743	100%

Table 3: Numbers of multi-labelled document in the mailing list data set.

Table 5 shows the mutual information, the p-values of Pearson's chi-square test of pairs of categories in the mailing list data set. From the table, we can see that some values of mutual information are clear not zero and some p-values show that the dependency between categories is significant (the smaller p-values indicate the stronger dependency between categories). Hence, the approach of combining single-labelled classifiers is insufficient for these data.

4.2 Methods and evaluation measures

For performance comparisons, we implemented two traditional methods and conducted performance evaluations using the same data corpora. The first method is the combination of multiple, independent single-labelled classifiers each of which employs the single-labelled maximum entropy model (which is equivalent to the penalized logistic regression), as described in Section 3.1. This method is denoted as "COMB" in our experiments. The second method is developed by stacking another layer of the penalized logistic regression on top of the first method, which adopts the idea of the approach described in [10]. We use the penalized logistic regression instead of SVM's because we want to use the same loss function for data classification model so that the results are more comparable. This method is denoted as "HF" in our experiments. Our proposed multi-labelled classification method based on the maximum entropy model is denoted as "MLME".

For a given document i, let $\mathbf{y}^{(i)}$ and $\hat{\mathbf{y}}^{(i)}$ be the true and the predicted label sets, respectively. We use the classification accuracy AC defined below as our performance metric.

$$\mathrm{AC} = \frac{\sum_{i=1}^{n} \delta(\mathbf{y}^{(i)}, \hat{\mathbf{y}}^{(i)})}{n}, \tag{16}$$

where n denotes the total number of documents in the test, $\delta(\mathbf{x}, \mathbf{y})$ is the delta function that equals one if $\mathbf{x} = \mathbf{y}$ for all dimensions and equals zero otherwise. AC computes the percentage of the documents whose predicted labels are exactly the same as their true labels.

Though the accuracy measures are compatible with the loss function of classification, which is considered as a smoothed version of $0-1$ loss, we are also interested in the practical goal of information retrieval. For multiple label data sets, we usually use micro-averaged F_1 measure [19],

$$F_1 = \frac{2rp}{r+p}, \tag{17}$$

where p and r are the precision rate and the recall rate computed globally over all binary decisions of all document-category pairs, respectively. Since the micro-averaged F_1 measure computes over all binary decisions, the partial correctness of labeling is credited.

4.3 Experimental results

The first experiment is on the Reuters-21578 document corpus. We used ten-fold cross validations to choose optimal regularization parameters for all the three methods. Table 6 shows the evaluation results using the optimal parameters on ten $9-1$ random splits. To compare the performance of different methods, we use one-sided Wilcoxon signed-rank test[18] which is a nonparametric paired test without assuming the underline distribution of the tested values. Here, the alternative hypothesis is whether multi-labelled maximum entropy method (MLME) has higher accuracies (or, F_1 measures). The p-values of one-sided Wilcoxon signed-rank test between the given experiment and multi-labelled maximum entropy method are shown in Table 6. Although the improvement of accuracies and F_1 measure is small, the improvement is significant (usually, if the p-value is smaller than 0.05 , the result is significant). The improvement is not much, partially because that the data set only contains 6.5% multi-labelled documents and the percentage of documents not belonging to the ten categories are relatively large, 32.4%.

The second experiment is on the mailing list data set. Table 7 shows the accuracy and F_1 measure for using their optimal parameters on ten 9-1 random splits. The accuracy and F_1 measure of the MLME method are better than those of the other two methods for every split, and the improvement is statistically significant.

4.4 The correlations among category labels

The intention of the proposed multi-labelled maximum entropy model is to include the correlations among categories into the model. Since the additional parameter R_{ij} is the coefficient of cate-

[2]The mailing lists are evolution-hackers@lists.ximian.com, freebsd-amd64@freebsd.org, freebsd-sparc64@freebsd.org, gnome-devel@gnome.org, image-sig@python.org, and public-esw@w3.org.

	acq	corn	crude	earn	grain	interest	money.fx	ship	trade	wheat
acq	0.17	0.0100	0.0120	0.1124	0.0240	0.0226	0.0305	0.0082	0.0195	0.0149
corn	1.3e-08	0.019	0.0060	0.0193	0.4398	0.0056	0.0048	0.0000	0.0034	0.1269
crude	6.9e-13	0.0093	0.041	0.0255	0.0070	0.0070	0.0101	0.0284	0.0018	0.0029
earn	3.5e-175	1.4e-18	2.5e-31	0.3	0.0416	0.0343	0.0500	0.0192	0.0362	0.0222
grain	7e-21	0	0.00018	9.2e-44	0.045	0.0100	0.0111	0.0106	0.0018	0.4967
interest	1e-17	0.015	0.0005	4.9e-35	6.6e-05	0.036	0.1123	0.0060	0.0018	0.0063
money.fx	1e-26	0.0048	4.9e-06	1.3e-54	1.1e-06	3.4e-182	0.056	0.0076	0.0014	0.0080
ship	6.3e-08	0.91	1.2e-30	1.6e-19	1.1e-10	0.011	0.00097	0.021	0.0024	0.0007
trade	3.5e-17	0.033	0.045	3e-37	0.037	0.052	0.023	0.058	0.038	0.0011
wheat	3.7e-11	1.6e-170	0.032	1.3e-21	0	0.0077	0.00059	0.29	0.19	0.022

Table 4: Some facts of data set Reuters-21578. The normalized symmetric mutual information values between categories are shown in the upper triangle. The p-values of Pearson's chi-square test (χ^2) for pairs of categories are shown in the lower triangle. The numbers in the diagonal are the proportions of categories.

	ASK	DEL	INFO	NEGCOM	POSCOM	RAISE	REP	SP
ASK	0.31	0.0020	0.0087	0.0002	0.0004	0.0768	0.0007	0.1125
DEL	0.037	0.022	0.0044	0.0000	0.0000	0.0027	0.0005	0.0082
INFO	3e-05	0.36	0.02	0.0031	0.0031	0.0032	0.0012	0.0170
NEGCOM	0.54	0.63	0.64	0.013	0.0034	0.0074	0.0079	0.0007
POSCOM	0.28	0.99	0.18	0.062	0.055	0.0257	0.0036	0.0015
RAISE	5.1e-78	0.012	0.0053	0.0017	3.9e-12	0.22	0.0513	0.4194
REP	0.11	0.41	0.23	0.013	0.01	3.5e-46	0.13	0.0943
SP	4e-120	1.4e-06	5.3e-12	0.22	0.03	0	1.4e-83	0.61

Table 5: Some facts of the mailing list data set. The normalized symmetric mutual information values between categories are shown in the upper triangle. The p-values of Pearson's chi-square test (χ^2) for pairs of categories are shown in the lower triangle. The numbers in the diagonal are the proportions of categories.

gory i and category j in the model, we expect that R_{ij} is somewhat related the correlation between category i and category j. Figures 1 and 2 plot correlations among categories and corresponding parameters of R from one of the experiment runs. The figures clearly show that the relation between the correlation and parameter R is significant and a pair with large correlation usually has a larger R parameter, especially when the correlations are far from zero.

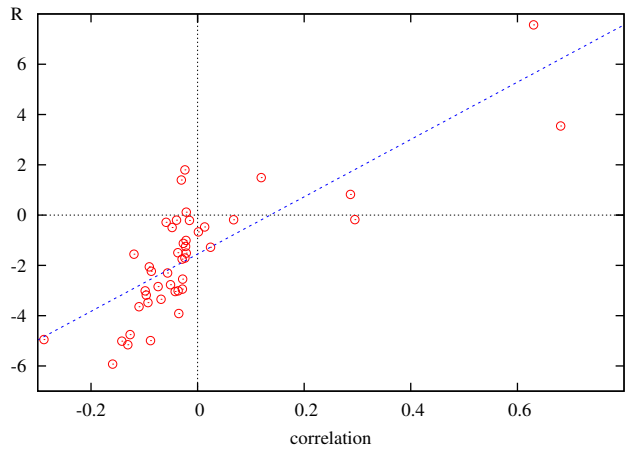


Figure 1: The correlations among categories and their corresponding parameters of R in the Reuters-21578 data set. The dotted line is the linear regression of the data points, which indicates the trend of relation between the correlations and R.

The R parameters enforce the correlations among categories in the model. These figures confirm our assumption and indicate that the correlation terms (strictly speaking the second order moments of labels) are important in these multi-labelled classification problems.

5. CONCLUDING REMARKS

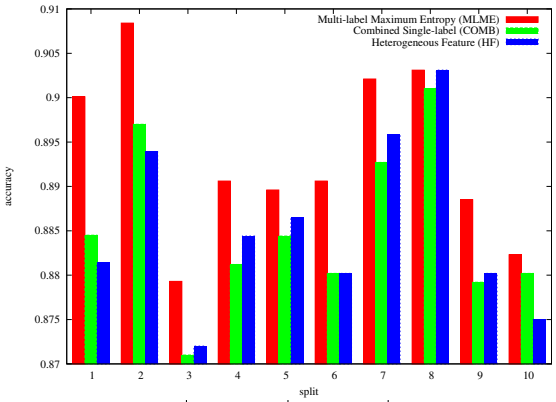
In this paper, we propose a maximum entropy method for multi-labelled classification, in which the correlations among category labels are explicitly considered in the model. The experimental results show that multi-labelled classification is beneficial in the model considering the correlation between classes, especially when the correlation is relatively strong. By examining the parameters of the model, the experiments confirm our assumption that the correlation terms are important in multi-labelled classification tasks.

One drawback of this method is in computing the term Z of Eq. (13). One possible solution is to use a stochastic approach. Another possible solution is to approximate $Z(\mathbf{x})$ with the sum of several important $q(\mathbf{y}|\mathbf{x})$.

During the simplification of the model, we assume that estimate errors are independent from each other. We do not know how large the impact is when this assumption does not hold. The future work may also involve the investigation of correlations among estimate errors.

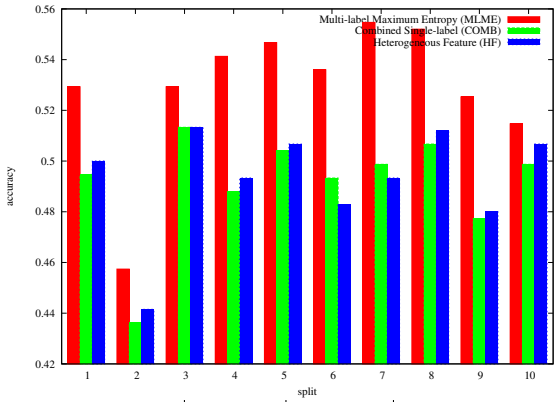
Acknowledgement

We would like to thank Mei Han, Tao Li and anonymous reviewers for useful comments and discussion. We thank the PETSc and TAO team of Argonne National Lab for their work.



	MLME	COMB	HF
Accuracy AC			
average	**0.8935**	0.8851	0.8852
p-value		0.0029	0.0045
test set	**0.8857**	0.8742	0.8724
Micro-averaged F_1			
average	**0.9155**	0.9105	0.9106
p-value		0.0068	0.0029
test set	**0.9180**	0.9104	0.9094

Table 6: The accuracies and F_1 measures of experiments on the Reuters-21578 dataset and their one-sided Wilcoxon signed-rank test vs MLME.



	MLME	COMB	HF
Accuracy AC			
average	**0.5287**	0.4911	0.4929
p-value		0.0029	0.0029
Micro-averaged F_1			
average	**0.6808**	0.6603	0.6659
p-value		0.00098	0.0068

Table 7: The accuracies and F_1 measures of experiments on the mailing list data set and their one-sided Wilcoxon signed-rank test vs MLME.

Appendix

Here are some details of how we derive Eq. (13). For parameters η_i and ϕ_{il} in Eq. (10) and θ_{ij} in Eq. (11), we regularize them to avoid the extreme results during estimating the parameters of the model. Assuming the joint probability of estimate errors should be reasonably large, say greater than a small number ϵ, we write

$$\Pr(\eta_i, \theta_{ij}, \phi_{il}) \geq \epsilon. \tag{18}$$

To simplify this constraint of Eq. (18), we assume that those estimate errors are independent to each other. Hence, we can rewrite this constraint in logarithm format as

$$\begin{aligned} \sum_i -\log \Pr(\eta_i) + \sum_{ij} -\log \Pr(\theta_{ij}) \\ + \sum_{il} -\log \Pr(\phi_{il}) \leq -\log \epsilon. \end{aligned} \tag{19}$$

According to the central limit theorem, the estimation errors follow normal distribution. Let $\eta_i \sim \mathcal{N}(0, \sigma_\eta^2/n)$, $\theta_{ij} \sim \mathcal{N}(0, \sigma_\theta^2/n)$ and $\phi_{il} \sim \mathcal{N}(0, \sigma_\phi^2/n)$, where n is the number of data points. The constraint can be simplified as

$$\sum_i \frac{\eta_i^2}{2\sigma_\eta^2/n} + \sum_{i<j} \frac{\theta_{ij}^2}{2\sigma_\theta^2/n} + \sum_{i,l} \frac{\phi_{il}^2}{2\sigma_\phi^2/n} \leq C, \tag{20}$$

where C is a constant derived from ϵ, σ's and n.

The Lagrangian of Eq. (3) subject to Eq. (10, 11, 20) and $\sum_{\mathbf{y}} q(\mathbf{y}|\mathbf{x}) = 1$ is:

$$\begin{aligned} \mathcal{L}(q(\mathbf{y}|\mathbf{x}), \eta, \theta, \phi, \underline{\ }R, W, \gamma, \zeta(\mathbf{x})) &= \langle \log q(\mathbf{y}|\mathbf{x}) \rangle_Q \\ &+ \sum_i b_i(\langle y_i \rangle_{\tilde{P}} + \eta_i - \langle y_i \rangle_Q) \\ &+ \sum_{i<j} R_{ij}(\langle y_i y_j \rangle_{\tilde{P}} + \theta_{ij} - \langle y_i y_j \rangle_Q) \\ &+ \sum_{i,l} W_{il}(\langle y_i x_l \rangle_{\tilde{P}} + \phi_{il} - \langle y_i x_l \rangle_Q) \\ &+ \gamma(\sum_i \frac{\eta_i^2}{2\sigma_\eta^2/n} + \sum_{i<j} \frac{\theta_{ij}^2}{2\sigma_\theta^2/n} + \sum_{i,l} \frac{\phi_{il}^2}{2\sigma_\phi^2/n} - C) \\ &+ \sum_{\mathbf{x}} \zeta(\mathbf{x})(1 - \sum_{\mathbf{y}} q(\mathbf{y}|\mathbf{x})), \end{aligned} \tag{21}$$

where $\mathbf{b}$, R (strict upper triangle matrix), W, γ (≥ 0), and ζ are the Lagrangian multipliers.

The Karush-Kuhn-Tucker (KKT) conditions require the derivatives of the Lagrangian with respect to its parameters must be zeros to maximize $\mathcal{L}$. Therefore, we have:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial q(\mathbf{y}|\mathbf{x})} &= \tilde{p}(\mathbf{x})[\log q(\mathbf{y}|\mathbf{x}) + 1 - \mathbf{y}^\top(\mathbf{b} + R\mathbf{y} + W\mathbf{x})] \\ &\quad - \zeta(\mathbf{x}) = 0, \\ \frac{\partial \mathcal{L}}{\partial \eta_i} &= b_i + n\gamma \frac{\eta_i}{\sigma_\eta^2} = 0, \\ \frac{\partial \mathcal{L}}{\partial \theta_{ij}} &= R_{ij} + n\gamma \frac{\theta_{ij}}{\sigma_\theta^2} = 0, \\ \frac{\partial \mathcal{L}}{\partial \phi_{ij}} &= W_{ik} + n\gamma \frac{\phi_{ik}}{\sigma_\phi^2} = 0. \end{aligned} \tag{22}$$

When γ is zero, the problem is trivial. Now we assume that $\gamma > 0$. It allows us to express q, η, θ and ϕ as functions of $\mathbf{b}$, R,

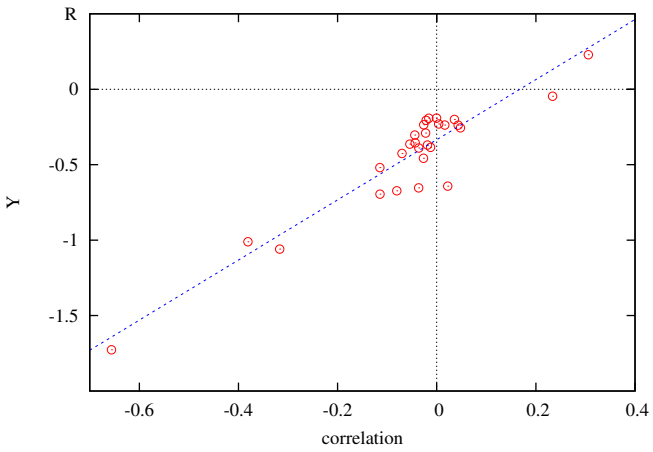


Figure 2: The correlations among categories and their corresponding parameters of R in the mailing list data set. The dotted line is the linear regression of the data points, which indicates the trend of relation between the correlations and R.

W, and γ:

$$\hat{q}(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp(\mathbf{y}^\top(\mathbf{b} + R\mathbf{y} + W\mathbf{x})),$$
$$\hat{\eta}_i = -\frac{\sigma_\eta^2}{n\gamma} b_i, \quad \hat{\theta}_{ij} = -\frac{\sigma_\theta^2}{n\gamma} R_{ij}, \quad \hat{\phi}_{ik} = -\frac{\sigma_\phi^2}{n\gamma} W_{ik}, \tag{23}$$

where the partition function, $Z(\mathbf{x}) = \sum_{\mathbf{y}} \exp(\mathbf{y}^\top(\mathbf{b} + R\mathbf{y} + W\mathbf{x}))$. By plugging Eq. (23) into Eq. (21), we have Eq. (13).

6. REFERENCES

[1] Benson, S. J., McInnes, L. C., Moré, J., & Sarich, J. (2004). *TAO user manual (revision 1.7)* (Technical Report ANL/MCS-TM-242). Mathematics and Computer Science Division, Argonne National Laboratory. http://www.mcs.anl.gov/tao.

[2] Cai, L., & Hofmann, T. (2004). Hierarchical document categorization with support vector machines. *CIKM '04: Proceedings of the Thirteenth ACM conference on Information and knowledge management* (pp. 78–87). Washington, D.C., USA: ACM Press.

[3] Chen, S. F., & Rosenfeld, R. (1999). *A Gaussian prior for smoothing maximum entropy models* (Technical Report CMU-CS-99-108). School of Computer Science Carnegie Mellon University.

[4] Clare, A., & King, R. D. (2001). Knowledge discovery in multi-label phenotype data. *PKDD '01: Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 42–53). Springer-Verlag.

[5] Comite, F. D., Gilleron, R., & Tommasi, M. (2001). Learning multi-label alternating decision trees and applications. *Proceedings of CAP'01* (pp. 195–210).

[6] Crammer, K., & Singer, Y. (2002). A new family of online algorithms for category ranking. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 151–158). Tampere, Finland: ACM Press.

[7] Della Pietra, S., Della Pietra, V. J., & Lafferty, J. D. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*, 380–393.

[8] Elisseeff, A., & Weston, J. (2002). A kernel method for multi-labelled classification. *Advances in Neural Information Processing Systems 14* (pp. 681–687). Cambridge, MA: MIT Press.

[9] Gao, S., Wu, W., Lee, C.-H., & Chua, T.-S. (2004). A mfom learning approach to robust multiclass multi-label text categorization. *ICML '04: Twenty-first international conference on Machine learning*. Banff, Alberta, Canada: ACM Press.

[10] Godbole, S., & Sarawagi, S. (2004). Discriminative methods for multi-labeled classification. *PAKDD*.

[11] Har-Peled, S., Roth, D., & Zimak, D. Constraint classification for multiclass classification and ranking. In S. T. S. Becker and K. Obermayer (Eds.), *Advances in neural information processing systems 15*. MIT Press.

[12] Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, *106*, 620–630.

[13] Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. *Proc. of the sixth CoNLL*.

[14] McCallum, A. (1999). Multi-label text classification with a mixture model trained by EM. *AAAI'99 Workshop on Text Learning*.

[15] Nigam, K., Lafferty, J., & McCallum, A. (1999). Using maximum entropy for text classification. *IJCAI-99 Workshop on Machine Learning for Information Filtering* (pp. 61–67).

[16] Schapire, R. E., & Singer, Y. (2000). Boostexter: A boosting-based system for text categorization. *Machine Learning*, *39*, 135–168.

[17] Ueda, N., & Saito, K. Parametric mixture models for multi-labeled text. *Advances in Neural Information Processing Systems 15*. MIT Press.

[18] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, *1*, 80–93.

[19] Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. *Proceedings of the 22nd Annual International Conference on Research and Development in Information Retrieval (SIGIR'99)* (pp. 42–49). Berkley: ACM Press.

[20] Zhang, T., & Oles, F. J. (2001). Text categorization based on regularized linear classification methods. *Inf. Retr.*, *4*, 5–31.

[21] Zhu, J., & Hastie, T. (2003). Classification of gene microarrays by penalized logistic regression. *Biostatistics*.