Fast Multi-label Learning

Xiuwen Gong, **Dong Yuan** and **Wei Bao**
Faculty of Engineering, The University of Sydney
{xiuwen.gong, dong.yuan, wei.bao}@sydney.edu.au

Abstract

Embedding approaches have become one of the most pervasive techniques for multi-label classification. However, the training process of embedding methods usually involves a complex quadratic or semidefinite programming problem, or the model may even involve an NP-hard problem. Thus, such methods are prohibitive on large-scale applications. More importantly, much of the literature has already shown that the binary relevance (BR) method is usually good enough for some applications. Unfortunately, BR runs slowly due to its linear dependence on the size of the input data. The goal of this paper is to provide a simple method, yet with provable guarantees, which can achieve competitive performance without a complex training process. To achieve our goal, we provide a simple stochastic sketch strategy for multi-label classification and present theoretical results from both algorithmic and statistical learning perspectives. Our comprehensive empirical studies corroborate our theoretical findings and demonstrate the superiority of the proposed methods.

1 Introduction

Multi-label classification [Prabhu and Varma, 2014; Yen *et al.*, 2016; Liu *et al.*, 2019; Gong *et al.*, 2020], in which each instance can belong to multiple labels simultaneously, has significantly attracted the attention of researchers as a result of its wide range of applications, which range from document classification and automatic image annotation to video annotation. For example, when classifying documents, one may need to classify them into different groups, such as *Science*, *Finance* and *Sports*. In automatic image annotation, one needs to automatically predict relevant keywords, such as *beach*, *sky* and *tree*, to describe a natural scene image.

A popular strategy in multi-label learning is binary relevance (BR)[Tsoumakas *et al.*, 2010], which independently trains a linear regression model for each label independently. Recently, some sophisticated models are developed to improve the performance of BR. For example, embedding approaches [Hsu *et al.*, 2009; Chen and Lin, 2012; Yu *et al.*, 2014; Liu and Tsang, 2017; Liu *et al.*, 2017] have become popular techniques. Even though embedding methods improve the prediction performance of BR to some extent, their training process usually involves a complex quadratic or semidefinite programming problem, as in [Zhang and Schneider, 2012], or their model may involve an NP-hard problem, as in [Yu *et al.*, 2014] and [Bhatia *et al.*, 2015]. Thus, these kinds of methods are prohibitive on large-scale applications. Much of the literature, such as [Luaces *et al.*, 2012], [Madjarov *et al.*, 2012] and [Taha and Tiun, 2016], has already shown that BR with appropriate base learner is usually good enough for some applications, such as document classification [Taha and Tiun, 2016]. Unfortunately, BR runs slowly due to its linear dependence on the size of the input data. The question is how to overcome these computational obstacles yet obtain comparable results with BR.

To address the above problem, we provide a simple stochastic sketch strategy for multi-label classification. In particular, we carefully construct a small sketch of the full data set, and then use that sketch as a surrogate to perform fast optimization. This paper first introduces stochastic σ-subgaussian sketch, and then proposes the construction of a sketch matrix based on Walsh-Hadamard matrix to reduce the expensive matrix multiplications of σ-subgaussian sketch. From an algorithmic perspective, we provide provable guarantees that our proposed methods are approximately as good as the exact solution of BR. From a statistical learning perspective, we provide the generalization error bound of multi-label classification using our proposed stochastic sketch model.

Experiments on various real-world data sets demonstrate the superiority of the proposed methods. The results verify our theoretical findings. We organize this paper as follows. The second section introduces our proposed stochastic sketch for multi-label classification. The third section provides the provable guarantees for our algorithm from both algorithmic and statistical learning perspectives, and experimental results are presented in the fourth section. The last section provides our conclusions.

2 Stochastic Sketch for Multi-label Classification

Assume that $x^{(i)} \in \mathbb{R}^{p\times 1}$ is a real vector representing an input (instance), and $y^{(i)} \in \{0,1\}^{q\times 1}$ is a real vector representing the corresponding output ($i \in \{1 \dots n\}$). n denotes the number of training samples. The input matrix is $X \in \mathbb{R}^{n \times p}$ and the output matrix is $Y \in \{0,1\}^{n \times q}$. $\langle \cdot, \cdot \rangle$ and $\mathbf{I}_{n \times n}$ represent the inner product and the $n \times n$ identity matrix, respectively. We denote the transpose of the vector/matrix by the superscript $'$ and the logarithms to base 2 by log. Let $|| \cdot ||_2$ and $|| \cdot ||_F$ represent the l_2 norm and Frobenius norm, respectively. Let $V \in \mathbb{R}^{p \times q}$ be the regressors and $N(0,1)$ denote the standard Gaussian distribution.

A simple linear regression model for BR [Tsoumakas *et al.*, 2010] learns the matrix V through the following formulation:

$$\min_{V \in \mathbb{R}^{p \times q}} \frac{1}{2} ||XV - Y||_F^2 \tag{1}$$

Assuming that $n > p$ and $n > q$, the computational complexity for this problem is $\mathcal{O}(npq + np^2)$ [Golub and Loan, 1996]. The computational cost of an exact solution for problem 1 will be prohibitive on large-scale settings. To solve this problem, we construct a small sketch of the full data set by stochastic projection methods, and then use that sketch as a surrogate to perform fast optimization for problem 1. Specifically, we define a sketch matrix $S \in \mathbb{R}^{m \times n}$ and $S \neq 0$, where $m < n$ is the projection dimension and 0 is the zero matrix with all the zero entries. The input matrix X and output matrix Y are approximated by their sketched matrix SX and SY, respectively. We aim to solve the following sketched problem of problem 1.

$$\min_{V \in \mathbb{R}^{p \times q}} \frac{1}{2} ||SXV - SY||_F^2 \tag{2}$$

Motivated by [Weinberger and Saul, 2009; Kulis, 2013; Bhatia *et al.*, 2015], we use a k-nearest neighbor (kNN) classifier in the embedding space for prediction, instead of using an expensive decoding process [Zhang and Schneider, 2012]. Next, we introduce two kinds of stochastic sketch methods.

2.1 Stochastic σ-Subgaussian Sketch

The entries of a sketch matrix can be simply defined as i.i.d random variables from certain distributions, such as Gaussian distribution and Bernoulli distribution. [Matousek, 2008] has already shown that each of these distributions is a special case of Subgaussian distribution, which is defined as follows:

Definition 1 (σ-Subgaussian)**.** *A row $s_i \in \mathbb{R}^n$ of the sketch matrix S is σ-Subgaussian, if it has zero mean and for any vector $\zeta \in \mathbb{R}^n$ and $\epsilon \geq 0$, we have*

$$P(|\langle s_i, \zeta \rangle| \geq \epsilon ||\zeta||_2) \leq 2e^{-\frac{n\epsilon^2}{2\sigma^2}}$$

Clearly, a vector with i.i.d standard Gaussian entries or Bernoulli entries is 1-Subgaussian. We refer any matrix $S \in \mathbb{R}^{m \times n}$ to a Subgaussian sketch if its rows are zero mean, 1-Subgaussian, and with the covariance matrix $cov(s_i) = \mathbf{I}_{n \times n}$. A Subgaussian sketch is straightforward to construct. However, given the Subgaussian sketch $S \in \mathbb{R}^{m \times n}$, the cost of computing SX and SY is $\mathcal{O}(npm)$ and $\mathcal{O}(nqm)$, respectively. Next, we introduce the following technique to reduce this time complexity.

2.2 Stochastic Walsh-Hadamard Sketch

Inspired by [Ailon and Chazelle, 2009], we propose to construct the sketch matrix based on Walsh-Hadamard matrix to reduce the expensive matrix multiplications of Subgaussian sketch. Formally, a stochastic Walsh-Hadamard sketch matrix $S \in \mathbb{R}^{m \times n}$ is obtained with i.i.d. rows of the form:

$$s_i = \sqrt{n} e_i H R, \quad i = 1, \cdots, m$$

where $\{e_1, \cdots, e_m\}$ is a random subset of m rows uniformly sampled from $\mathbf{I}_{n \times n}$, $R \in \mathbb{R}^{n \times n}$ is a random diagonal matrix whose entries are i.i.d. Rademacher variables and $H \in \mathbb{R}^{n \times n}$ constitutes a Walsh-Hadamard matrix defined as:

$$H_{ij} = (-1)^{\langle \mathbb{B}(i)-1, \mathbb{B}(j)-1 \rangle}, \quad i,j = 1, \cdots, n$$

where $\mathbb{B}(i)$ and $\mathbb{B}(j)$ represent the binary expression with τ-bit of i and j (assume $2^\tau = n$).

Then, we can employ fast Walsh-Hadamard transform [Fino and Algazi, 1976] to perform SX and SY in $\mathcal{O}(np \log m)$ and $\mathcal{O}(nq \log m)$.

3 Main Results

Since we address problem 2 rather than directly solving problem 1, which has great advantages for fast optimization, it is interesting to ask the question: what is the relationship between problem 2 and problem 1? Let V^* and $\hat{V}$ be the optimal solutions of problem 1 and problem 2. We define $f(V^*) = ||XV^* - Y||_F^2$ and $g(\hat{V}) = ||SX\hat{V} - SY||_F^2$. We will prove that we can choose an appropriate m such that the two optimal objectives $f(V^*)$ and $g(\hat{V})$ are approximately the same. This means that we can speed up the computation of problem 1, without sacrificing too much accuracy. Furthermore, we provide the generalization error bound of the multi-label classification problem using our proposed stochastic sketch model. To measure the quality of approximation, we first define the δ-optimality approximation as follows:

Definition 2 (δ-Optimality Approximation)**.** *Given $\delta \in (0,1)$, $\hat{V}$ is a δ-optimality approximation solution, if*

$$(1-\delta) f(V^*) \leq g(\hat{V}) \leq (1+\delta) f(V^*)$$

According to the properties of Matrix norm, we have $g(\hat{V}) \leq ||S||_F f(\hat{V})$, so $g(\hat{V})$ is proportional to $f(\hat{V})$. Therefore, the closeness of $g(\hat{V})$ and $f(V^*)$ implies the closeness of $f(\hat{V})$ and $f(V^*)$.

3.1 σ-Subgaussian Sketch Guarantee

We first introduce the tangent cone, which is used by [Rockafellar and Wets, 2004]:

Definition 3 (Tangent Cone)**.** *Given a set $\mathcal{C} \subseteq \mathbb{R}^p$ and $x^* \in \mathcal{C}$, the tangent cone of $\mathcal{C}$ at x^* is defined as $\mathcal{K} = clconv\{r \in \mathbb{R}^p | r = t(x - x^*)$ for some $t \geq 0$ and $x \in \mathcal{C}\}$, where clconv denotes the closed convex hull.*

The tangent cone arises naturally in the convex optimality conditions: any $r \in \mathcal{K}$ defines a feasible direction at the optimal x^*, and optimality means that it is impossible to decrease the objective function by moving in directions belonging to the tangent cone. Then, we introduce the Gaussian width, which is an important complexity measure used by [Gordon, 1985]:

Definition 4 (Gaussian Width)**.** *Given a closed set $\mathcal{Y} \subseteq \mathbb{R}^n$, the Gaussian width of $\mathcal{Y}$, denoted by $\omega(\mathcal{Y})$, is defined as:*

$$\omega(\mathcal{Y}) = \mathbb{E}_g[\sup_{z \in \mathcal{Y}} |\langle g, z \rangle|]$$

where $g \sim N(0, \mathbf{I}_{n \times n})$.

This complexity measure plays an important role in learning theory and statistics [Koltchinskii and Panchenko, 2000]. Let $\mathbb{S}^{n-1} = \{z \in \mathbb{R}^n | ||z||_2 = 1\}$ be the Euclidean sphere. $X\mathcal{K}$ represents the linearly transformed cone: $\{Xr \in \mathbb{R}^n | r \in \mathcal{K}\}$, and we use Gaussian width to measure the width of the intersection of $X\mathcal{K}$ and $\mathbb{S}^{n-1}$. This paper defines $\mathcal{Y} = X\mathcal{K} \cap \mathbb{S}^{n-1}$. We state the following theorem for guaranteeing the σ-Subgaussian sketch:

Theorem 1. *Let $S \in \mathbb{R}^{m \times n}$ be a stochastic σ-Subgaussian sketch matrix, c_1 and c_2 be universal constants. Given any $\delta \in (0, 1)$ and $m = \mathcal{O}((\frac{c_1}{\delta})^2 \omega^2(\mathcal{Y}))$, then with probability at least $1 - 6qe^{-\frac{c_2 m \delta^2}{\sigma^4}}$, $\hat{V}$ is a δ-optimality approximation solution.*

The proof sketch of this theorem can be found in the Appendix.

Remark. Theorem 1 guarantees that the stochastic σ-Subgaussian sketch method is able to construct a small sketch of the full data set for the fast optimization of problem 1, while preserving the δ-optimality of the solution.

3.2 Walsh-Hadamard Sketch Guarantee

We generalize the concept of Gaussian width to two additional measures, S-Gaussian width and Rademacher width:

Definition 5 (S-Gaussian Width)**.** *Given a closed set $\mathcal{Y} \subseteq \mathbb{R}^n$ and a stochastic sketch matrix $S \in \mathbb{R}^{m \times n}$, the S-Gaussian width of $\mathcal{Y}$, denoted by $\omega_S(\mathcal{Y})$, is defined as:*

$$\omega_S(\mathcal{Y}) = \mathbb{E}_{g,S}[\sup_{z \in \mathcal{Y}} |\langle g, \frac{Sz}{\sqrt{m}} \rangle|]$$

where $g \sim N(0, \mathbf{I}_{m \times m})$.

Definition 6 (Rademacher Width)**.** *Given a closed set $\mathcal{Y} \subseteq \mathbb{R}^n$, the Rademacher width of $\mathcal{Y}$, denoted by $\Upsilon(\mathcal{Y})$, is defined as:*

$$\Upsilon(\mathcal{Y}) = \mathbb{E}_{\varpi}[\sup_{z \in \mathcal{Y}} |\langle \varpi, z \rangle|]$$

where $\varpi \in \{\pm 1\}^n$ is an i.i.d. vector of Rademacher variables.

Next, we still define $\mathcal{Y} = X\mathcal{K} \cap \mathbb{S}^{n-1}$ and state the following theorem for guaranteeing the Walsh-Hadamard sketch:

Theorem 2. *Let $S \in \mathbb{R}^{m \times n}$ be a stochastic Walsh-Hadamard sketch matrix, c_1, c_2 and c_3 be universal constants. Given any $\delta \in (0, 1)$ and $m = \mathcal{O}((\frac{c_1}{\delta})^2 (\Upsilon(\mathcal{Y}) + \sqrt{6log(n)})^2 \omega_S^2(\mathcal{Y}))$, then with probability at least $1 - 6q\big(\frac{c_2}{(mn)^2} + c_2 e^{-\frac{c_3 m \delta^2}{\Upsilon(\mathcal{Y})^2 + log(nm)}}\big)$, $\hat{V}$ is a δ-optimality approximation solution.*

Remark. An additional term $(\Upsilon(\mathcal{Y}) + \sqrt{6log(n)})^2$ appears in the sketch size, so the required sketch size for the Walsh-Hadamard sketch is larger than that required for the σ-Subgaussian sketch. However, the potentially larger sketch size is offset by the much lower cost of matrix multiplications via the stochastic Walsh-Hadamard sketch matrix. Theorem 2 guarantees that the stochastic Walsh-Hadamard sketch method is also able to construct a small sketch of the full data set for the fast optimization of problem 1, while preserving the δ-optimality of the solution.

3.3 Generalization Error Bound

This subsection provides the generalization error bound of the multi-label classification problem using our proposed two stochastic sketch models. Because our results can be applied to two models, we simply call our stochastic sketch models SS. Assume our model is characterized by a distribution $\mathcal{D}$ on the space of inputs and labels $\mathcal{X} \times \{0, 1\}^q$, where $\mathcal{X} \subseteq \mathbb{R}^p$. Let a sample $\{(x^{(j)}, y^{(j)})\}$ be drawn i.i.d. from the distribution $\mathcal{D}$, where $y^{(j)} \in \{0, 1\}^q$ ($j \in \{1, \dots, n\}$) are the ground truth label vectors. Assume n samples $D = \{(x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)})\}$ are drawn i.i.d. n times from the distribution $\mathcal{D}$, which is denoted by $D \sim \mathcal{D}^n$. For two inputs $x^{(z)}, x^{(j)}$ in $\mathcal{X}$, we define $d(x^{(z)}, x^{(j)}) = ||x^{(z)} - x^{(j)}||_2$ as the Euclidean metric in the original input space and $d_{pro}(x^{(z)}, x^{(j)}) = \|\hat{V}' x^{(z)} - \hat{V}' x^{(j)}\|_2$ as the metric in the embedding input space. Let $h^D_{knn_i}(x)$ represent the prediction of the i-th label for input x using our model SS-kNN, which is trained on D. The performance of SS-kNN: $(h^D_{knn_1}(\cdot), \cdots, h^D_{knn_q}(\cdot)) : \mathcal{X} \to \{0, 1\}^q$ is then measured in terms of its generalization error, which is its expected loss on a new example (x, y) drawn according to $\mathcal{D}$:

$$E_{D \sim \mathcal{D}^n, (x,y) \sim \mathcal{D}} \Big(\sum_{i=1}^{q} \ell(y_i, h^D_{knn_i}(x)) \Big) \tag{3}$$

where y_i means the i-th label and $\ell(y_i, h^D_{knn_i}(x))$ represents the loss function for the i-th label. We define the loss function as follows for the analysis.

$$\ell(y_i, h^D_{knn_i}(x)) = P(y_i \neq h^D_{knn_i}(x)) \tag{4}$$

For the i-th label, we define the function as follows:

$$\nu^i_j(x) = P(y_i = j | x), j \in \{0, 1\}. \tag{5}$$

The Bayes optimal classifier b^* for the i-th label is defined as

$$b^*_i(x) = \arg \max_{j \in \{0,1\}} \nu^i_j(x) \tag{6}$$

Before deriving our results, we first present several important definitions and theorems.

Definition 7 (Covering Numbers, [Shawe-Taylor *et al.*, 1998])**.** *Let $(\mathcal{X}, d)$ be a metric space, A be a subset of $\mathcal{X}$ and $\varepsilon > 0$. A set $B \subseteq X$ is an ε-cover for A, if for every $a \in A$, there exists $b \in B$ such that $d(a, b) < \varepsilon$. The ε-covering number of A, $\mathcal{N}(\varepsilon, A, d)$, is the minimal cardinality of an ε-cover for A (if there is no such finite cover then it is defined as ∞).*

					SS+GAU			SS+WH		
DATA SET	BR+LIB	BR+kNN	FASTXML	SLEEC	$m = 256$	$m = 512$	$m = 1024$	$m = 256$	$m = 512$	$m = 1024$
COREL5K	0.0098	0.0095	0.0093	0.0094	0.0095	0.0095	0.0094	0.0103	0.0102	0.0099
NUS(VLAD)	0.0211	0.0213	0.0209	0.0207	0.0221	0.0218	0.0216	0.0230	0.0225	0.0218
NUS(BOW)	0.0215	0.0220	0.0216	0.0213	0.0227	0.0223	0.0222	0.0229	0.0226	0.0223
RCV1X	0.0017	0.0019	0.0019	0.0018	0.00189	0.00188	0.00187	0.00199	0.00195	0.00192

Table 1: The results of Hamming Loss on the various data sets.

					SS+GAU			SS+WH		
DATA SET	BR+LIB	BR+kNN	FASTXML	SLEEC	$m = 256$	$m = 512$	$m = 1024$	$m = 256$	$m = 512$	$m = 1024$
COREL5K	0.1150	0.0930	0.0530	0.0824	0.0475	0.0446	0.0659	0.0539	0.0817	0.0902
NUS(VLAD)	0.1247	0.1547	0.1118	0.1578	0.1099	0.1310	0.1460	0.1001	0.1289	0.1443
NUS(BOW)	0.0984	0.1012	0.0892	0.1122	0.0896	0.0932	0.0952	0.0882	0.0903	0.0920
RCV1X	0.2950	0.2894	0.2367	0.2801	0.2063	0.2767	0.2813	0.2173	0.2621	0.2796

Table 2: The results of Example-F1 on the various data sets.

Definition 8 (Doubling Dimension, [Krauthgamer and Lee, 2004]). *Let $(\mathcal{X}, d)$ be a metric space, and let $\bar{\lambda}$ be the smallest value such that every ball in $\mathcal{X}$ can be covered by $\bar{\lambda}$ balls of half the radius. The doubling dimension of $\mathcal{X}$ is defined as : $ddim(\mathcal{X}) = \log_2(\bar{\lambda})$.*

Theorem 3 ([Krauthgamer and Lee, 2004]). *Let $(\mathcal{X}, d)$ be a metric space. The diameter of $\mathcal{X}$ is defined as $diam(\mathcal{X}) = \sup\limits_{x,x' \in \mathcal{X}} d(x, x')$. The ε-covering number of $\mathcal{X}$, $\mathcal{N}(\varepsilon, \mathcal{X}, d)$, is bounded by:*

$$\mathcal{N}(\varepsilon, \mathcal{X}, d) \leq \Big(\frac{2diam(\mathcal{X})}{\varepsilon}\Big)^{ddim(\mathcal{X})} \tag{7}$$

We provide the following generalization error bound for SS-1NN:

Theorem 4. *Given a metric space $(\mathcal{X}, d_{pro})$, assume function $\nu^i : \mathcal{X} \to [0, 1]$ is Lipschitz with constant L with respect to the sup-norm for each label. Suppose $\mathcal{X}$ has a finite doubling dimension: $ddim(\mathcal{X}) = \mathbb{D} < \infty$ and $diam(\mathcal{X}) = 1$. Let $D = \{(x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)})\}$ and (x, y) be drawn i.i.d. from the distribution $\mathcal{D}$. Then, we have*

$$\begin{aligned}&E_{D\sim\mathcal{D}^n,(x,y)\sim\mathcal{D}}\Big(\sum_{i=1}^{q} P(y_i \neq h^D_{1nn_i}(x))\Big)\\ &\leq \sum_{i=1}^{q} 2P(b^*_i(x) \neq y_i) + \frac{3qL||\hat{V}||_F}{n^{1/(\mathbb{D}+1)}}\end{aligned} \tag{8}$$

Inspired by Theorem 19.5 in [Shalev-Shwartz and Ben-David, 2014], we derive the following lemma for SS-kNN:

Lemma 1. *Given metric space $(\mathcal{X}, d_{pro})$, assume function $\nu^i : \mathcal{X} \to \{0, 1\}$ is Lipschitz with constant L with respect to the sup-norm for each label. Suppose $\mathcal{X}$ has a finite doubling dimension: $ddim(\mathcal{X}) = \mathbb{D} < \infty$ and $diam(\mathcal{X}) = 1$. Let $D = \{(x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)})\}$ and (x, y) be drawn i.i.d. from the distribution $\mathcal{D}$. Then, we have*

$$\begin{aligned}&E_{D\sim\mathcal{D}^n,(x,y)\sim\mathcal{D}}\Big(\sum_{i=1}^{q} P(y_i \neq h^D_{knn_i}(x))\Big)\\ &\leq \sum_{i=1}^{q}(1 + \sqrt{8/k})P(b^*_i(x) \neq y_i) + \frac{q(6L||\hat{V}||_F + k)}{n^{1/(\mathbb{D}+1)}}\end{aligned} \tag{9}$$

The following corollary reveals important statistical properties of SS-1NN and SS-kNN.

Corollary 1. *As n goes to infinity, the error of the SS-1NN and SS-kNN converges to the sum of twice the Bayes error and $1 + \sqrt{8/k}$ times Bayes error over the labels, respectively.*

4 Experiment

4.1 Data Sets and Baselines

We abbreviate our proposed stochastic σ-Subgaussian sketch and stochastic Walsh-Hadamard sketch to SS+GAU and SS+WH, respectively. In the experiment, we set the entries in the σ-Subgaussian sketch matrix as i.i.d standard Gaussian entries. This section evaluates the performance of the proposed methods on four data sets: corel5k, nus(vlad), nus(bow) and rcv1x. The statistics of these data sets are presented in website[1]. We compare SS+GAU and SS+WH with several state-of-the-art methods, as follows.

- BR [Tsoumakas *et al.*, 2010]: We implement two base classifiers for BR. The first uses linear classification/regression package LIBLINEAR [Fan *et al.*, 2008] with l_2-regularized square hinge loss as the base classifier. We simply call this baseline BR+LIB. The second uses kNN as the base classifier. We simply call this baseline BR+kNN and count the kNN search time as the training time.
- FastXML [Prabhu and Varma, 2014]: An advanced tree-based multi-label classifier.
- SLEEC [Bhatia *et al.*, 2015]: A state-of-the-art embedding method, which is based on sparse local embeddings

[1] http://mulan.sourceforge.net

					SS+GAU			SS+WH		
DATA SET	BR+LIB	BR+kNN	FASTXML	SLEEC	$m = 256$	$m = 512$	$m = 1024$	$m = 256$	$m = 512$	$m = 1024$
COREL5K	7.198	0.678	4.941	736.670	0.196	0.218	0.366	0.119	0.197	0.239
NUS(VLAD)	222.21	179.04	715.86	9723.49	25.29	51.68	93.97	11.87	20.22	33.04
NUS(BOW)	511.83	351.64	1162.53	11391.54	52.05	72.65	120.37	25.41	34.32	48.85
RCV1X	22607.53	353.42	1116.05	78441.93	72.53	114.55	144.17	48.88	55.94	72.22

Table 3: The training time (in second) on the various data sets.

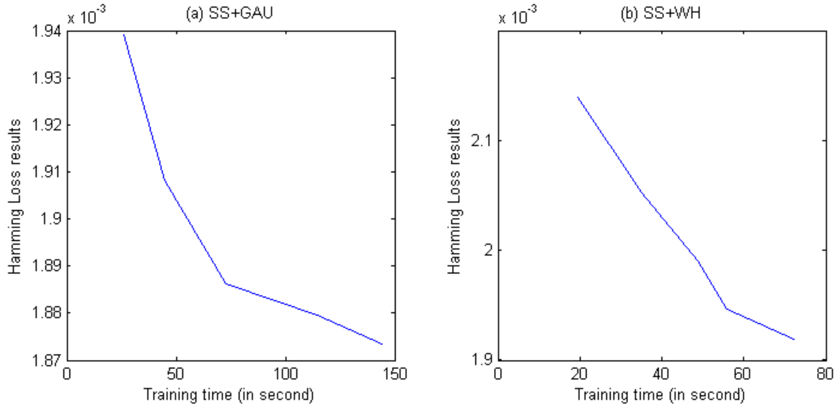

Figure 1: Experiment results of SS+GAU and SS+WH on rcv1x data set.

for large-scale multi-label classification. We use solvers of FastXML and SLEEC provided by the respective authors with default parameters.

Following the similar settings in [Zhang and Zhou, 2007] and [Bhatia *et al.*, 2015], we set $k = 10$ for the kNN search in all kNN based methods. The sketch size m is chosen in a range of $\{64, 128, 256, 512, 1024\}$. Following [Chen and Lin, 2012], [Zhang and Schneider, 2012] and [Guo and Schuurmans, 2013], we consider the Hamming Loss and Example-F1 measures to evaluate the prediction performance of all the methods. The smaller the value of the Hamming Loss, the better the performance, while the larger the value of Example-F1, the better the performance.

4.2 Results

Figure 1 shows that with the increasing sketch size, the training time of SS+GAU and SS+WH rise, while the prediction performance of SS+GAU and SS+WH becomes better. The results verify our theoretical analysis. The Hamming Loss, Example-F1 and training time comparisons of various methods on corel5k, nus(vlad), nus(bow) and rcv1x data sets are shown in Table 1, Table 2 and Table 3, respectively. From Tables 1, 2 and 3, we can see that:

- BR and SLEEC usually achieve better results, which is consistent with the empirical results in [Bhatia *et al.*, 2015] and [Taha and Tiun, 2016]. However, SLEEC is the slowest method compared to other baselines.
- Because we perform the optimization only on a small sketch of the full data set, our proposed methods are significantly faster than BR and state-of-the-art embedding approaches. Moreover, we can maintain competitive prediction performance by setting an appropriate sketch size. The empirical results illustrate our theoretical studies.

5 Conclusion

This paper carefully constructs stochastic σ-Subgaussian sketch and Walsh-Hadamard sketch for multi-label classification. From an algorithmic perspective, we show that we can obtain answers that are approximately as good as the exact answer for BR. From a statistical learning perspective, we also provide the generalization error bound of multi-label classification using our proposed stochastic sketch model. Lastly, our empirical studies corroborate our theoretical findings, and demonstrate the superiority of the proposed methods.

Acknowledgments

The author would like to thank all the anonymous reviewers for their useful comments.

A The Proof Sketch of Theorem 1

We first present the following lemma, which is derived from [Mendelson *et al.*, 2007].

Lemma 2. *Let $S \in \mathbb{R}^{m \times n}$ be a stochastic σ-Subgaussian sketch matrix. Then there are universal constants c_1 and c_2 such that for any subset $\mathcal{Y} \subseteq \mathbb{S}^{n-1}$, any $u \in \mathbb{S}^{n-1}$ and $\delta \in (0, 1)$, we have*

$$\sup_{z \in \mathcal{Y}} |z' \mathscr{S} z| \leq \frac{c_1}{\sqrt{m}} \omega(\mathcal{Y}) + \delta \tag{10}$$

with probability at least $1 - e^{-\frac{c_2 m \delta^2}{\sigma^4}}$, and we have

$$\sup_{z \in \mathcal{Y}} |z' \mathscr{S} u| \leq \frac{5c_1}{\sqrt{m}} \omega(\mathcal{Y}) + 3\delta \tag{11}$$

with probability at least $1 - 3e^{-\frac{c_2 m\delta^2}{\sigma^4}}$*, where* $\mathscr{S} = S'S - \mathbf{I}_{n\times n}$.

Proof. (of Theorem 1). We first show that $\hat{V}$ is a δ-optimality approximation solution from the right hand. Let $V_1^*, \cdots, V_q^*$, $\hat{V}_1, \cdots, \hat{V}_q$ and $Y_1, \cdots, Y_q$ be q columns of matrix V^*, $\hat{V}$ and Y, respectively. Then $||XV^* - Y||_F^2$ and $||SX\hat{V} - SY||_F^2$ can be decomposed to $||XV^* - Y||_F^2 = \sum_{i=1}^q ||XV_i^* - Y_i||_2^2$ and $||SX\hat{V} - SY||_F^2 = \sum_{i=1}^q ||SX\hat{V}_i - SY_i||_2^2$. Next, we study the relationship between $||XV_i^* - Y_i||_2^2$ and $||SX\hat{V}_i - SY_i||_2^2$. We define $M = \hat{V}_i - V_i^*$. According to Definition 3, we know that M belongs to the tangent cone of $\mathcal{C}$ at V_i^*.

Because $V_i^* \in \arg\min_{r\in\mathbb{R}^p} ||Xr - Y_i||_2^2$, we have $||XV_i^* - Y_i||_2^2 \le ||X\hat{V}_i - Y_i||_2^2 = ||XV_i^* - Y_i||_2^2 + 2\langle XV_i^* - Y_i, XM\rangle + ||XM||_2^2$. Consequently, we obtain:

$$2\langle XV_i^* - Y_i, XM\rangle + ||XM||_2^2 \ge 0 \tag{12}$$

As $\hat{V}_i \in \arg\min_{r\in\mathbb{R}^p} ||SXr - SY_i||_2^2$, we have $||SXV_i^* - SY_i||_2^2 \ge ||SX\hat{V}_i - SY_i||_2^2 = ||SXV_i^* - SY_i||_2^2 + 2\langle SXV_i^* - SY_i, SXM\rangle + ||SXM||_2^2$. Consequently, we obtain $||SXM||_2^2 \le -2\langle SXV_i^* - SY_i, SXM\rangle \le 2||SXV_i^* - SY_i||_2||SXM||_2$ and

$$||SXM||_2 \le 2||SXV_i^* - SY_i||_2 \tag{13}$$

We derive the following:

$$\begin{aligned}
&||SX\hat{V}_i - SY_i||_2^2 \\
=&||SXV_i^* - SY_i||_2^2 + ||SXM||_2^2 + 2\langle SXV_i^* - SY_i, SXM\rangle \\
=&||SXV_i^* - SY_i||_2^2 + ||XM||_2^2 + \langle XM, \mathscr{S}XM\rangle \\
&+ 2\langle XV_i^* - Y_i, \mathscr{S}XM\rangle + 2\langle XV_i^* - Y_i, XM\rangle
\end{aligned}$$

By using Lemma 2, with probability at least $1 - 4e^{-\frac{c_2 m\delta^2}{\sigma^4}}$, we have

$$\begin{aligned}
&||SX\hat{V}_i - SY_i||_2^2 \\
\le&||SXV_i^* - SY_i||_2^2 + ||XM||_2^2(1 + \frac{c_1}{\sqrt{m}}\omega(\mathcal{Y}) + \delta) \\
&+ 2||XV_i^* - Y_i||_2||XM||_2(1 + \frac{5c_1}{\sqrt{m}}\omega(\mathcal{Y}) + 3\delta)
\end{aligned}$$

where $\mathcal{Y} = X\mathcal{K} \cap \mathbb{S}^{n-1}$. Given $\gamma > 0$, we have $2||XV_i^* - Y_i||_2||XM||_2 \le \gamma||XV_i^* - Y_i||_2^2 + 1/\gamma||XM||_2^2$. For the sake of clarity, we define $\psi = 1 + \frac{5c_1}{\sqrt{m}}\omega(\mathcal{Y}) + 3\delta$ and $\varphi = 1 + \frac{c_1}{\sqrt{m}}\omega(\mathcal{Y}) + \delta$, and then substitute them to the above expression, with probability at least $1 - 4e^{-\frac{c_2 m\delta^2}{\sigma^4}}$, we have

$$\begin{aligned}
&||SX\hat{V}_i - SY_i||_2^2 \\
\le&||SXV_i^* - SY_i||_2^2 + \gamma\psi||XV_i^* - Y_i||_2^2 + (\frac{\psi}{\gamma} + \varphi)||XM||_2^2
\end{aligned} \tag{14}$$

Clearly, we have $\omega(\frac{XV_i^* - Y_i}{||XV_i^* - Y_i||_2}) \le \omega(\mathcal{Y})$. By using Lemma 2, with probability at least $1 - e^{-\frac{c_2 m\delta^2}{\sigma^4}}$, we have

$$\begin{aligned}
&||SXV_i^* - SY_i||_2^2 \\
=&||XV_i^* - Y_i||_2^2 + \langle XV_i^* - Y_i, \mathscr{S}(XV_i^* - Y_i)\rangle \\
\le&||XV_i^* - Y_i||_2^2(1 + \frac{c_1}{\sqrt{m}}\omega(\frac{XV_i^* - Y_i}{||XV_i^* - Y_i||_2}) + \delta) \\
\le&||XV_i^* - Y_i||_2^2\varphi
\end{aligned} \tag{15}$$

By using Lemma 2, with probability at least $1 - e^{-\frac{c_2 m\delta^2}{\sigma^4}}$, we have $||SXM||_2^2 = ||XM||_2^2 + \langle XM, \mathscr{S}XM\rangle \ge ||XM||_2^2(1 - \frac{c_1}{\sqrt{m}}\omega(\mathcal{Y}) - \delta) = ||XM||_2^2(2 - \varphi)$. By using Eq.(13), with probability at least $1 - e^{-\frac{c_2 m\delta^2}{\sigma^4}}$, we have

$$||XM||_2^2 \le \frac{||SXM||_2^2}{2 - \varphi} \le 4\frac{||SXV_i^* - SY_i||_2^2}{2 - \varphi} \tag{16}$$

Eq.(14), Eq.(15) and Eq.(16) imply that, with probability at least $1 - 6e^{-\frac{c_2 m\delta^2}{\sigma^4}}$, we have

$$\begin{aligned}
&||SX\hat{V}_i - SY_i||_2^2 \\
\le&(1 + 4\frac{\frac{\psi}{\gamma} + \varphi}{2 - \varphi})||SXV_i^* - SY_i||_2^2 + \gamma\psi||XV_i^* - Y_i||_2^2 \\
\le&(1 + 4\frac{\frac{\psi}{\gamma} + \varphi}{2 - \varphi})\varphi||XV_i^* - Y_i||_2^2 + \gamma\psi||XV_i^* - Y_i||_2^2 \\
\le&(\varphi - 4\frac{\psi}{\gamma} - 4\varphi + \gamma\psi)||XV_i^* - Y_i||_2^2
\end{aligned} \tag{17}$$

By setting $\gamma = 4$, with probability at least $1 - 6e^{-\frac{c_2 m\delta^2}{\sigma^4}}$, we have

$$\begin{aligned}
&||SX\hat{V}_i - SY_i||_2^2 \\
\le&(3\psi - 3\varphi)||XV_i^* - Y_i||_2^2 \\
=&(\frac{12c_1}{\sqrt{m}}\omega(\mathcal{Y}) + 6\delta)||XV_i^* - Y_i||_2^2
\end{aligned} \tag{18}$$

Eq.(18) implies that, with probability at least $1 - 6qe^{-\frac{c_2 m\delta^2}{\sigma^4}}$, we have

$$||SX\hat{V} - SY||_F^2 \le (\frac{12c_1}{\sqrt{m}}\omega(\mathcal{Y}) + 6\delta)||XV^* - Y||_F^2 \tag{19}$$

□

References

[Ailon and Chazelle, 2009] Nir Ailon and Bernard Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009.

[Bhatia *et al.*, 2015] Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. Sparse local embeddings for extreme multi-label classification. In *NIPS*, pages 730–738, 2015.

[Chen and Lin, 2012] Yao-Nan Chen and Hsuan-Tien Lin. Feature-aware label space dimension reduction for multi-label classification. In *NIPS*, pages 1538–1546, 2012.

[Fan *et al.*, 2008] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[Fino and Algazi, 1976] Bernard J. Fino and V. Ralph Algazi. Unified matrix treatment of the fast Walsh-Hadamard transform. *IEEE Transactions on Computers*, 25(11):1142–1146, 1976.

[Golub and Loan, 1996] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.

[Gong *et al.*, 2020] Xiuwen Gong, Dong Yuan, and Wei Bao. Online metric learning for multi-label classification. In *AAAI*, pages 4012–4019, 2020.

[Gordon, 1985] Yehoram Gordon. Some inequalities for Gaussian processes and applications. *Israel J. Math*, 50:109–110, 1985.

[Guo and Schuurmans, 2013] Yuhong Guo and Dale Schuurmans. Multi-label classification with output kernels. In *ECML/PKDD*, pages 417–432, 2013.

[Hsu *et al.*, 2009] Daniel Hsu, Sham Kakade, John Langford, and Tong Zhang. Multi-label prediction via compressed sensing. In *Advances in Neural Information Processing Systems 22*, pages 772–780, 2009.

[Koltchinskii and Panchenko, 2000] Vladimir Koltchinskii and Dmitriy Panchenko. *Rademacher Processes and Bounding the Risk of Function Learning*. Springer-Verlag, 2000.

[Krauthgamer and Lee, 2004] Robert Krauthgamer and James R. Lee. Navigating nets: Simple algorithms for proximity search. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 798–807, 2004.

[Kulis, 2013] Brian Kulis. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2013.

[Liu and Tsang, 2017] Weiwei Liu and Ivor W. Tsang. Making decision trees feasible in ultrahigh feature and label dimensions. *Journal of Machine Learning Research*, 18:81:1–81:36, 2017.

[Liu *et al.*, 2017] Weiwei Liu, Ivor W. Tsang, and Klaus-Robert Müller. An easy-to-hard learning paradigm for multiple classes and multiple labels. *Journal of Machine Learning Research*, 18:94:1–94:38, 2017.

[Liu *et al.*, 2019] Weiwei Liu, Donna Xu, Ivor W. Tsang, and Wenjie Zhang. Metric learning for multi-output tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):408–422, 2019.

[Luaces *et al.*, 2012] Oscar Luaces, Jorge Díez, José Barranquero, Juan José del Coz, and Antonio Bahamonde. Binary relevance efficacy for multilabel classification. *Progress in AI*, 1(4):303–313, 2012.

[Madjarov *et al.*, 2012] Gjorgji Madjarov, Dragi Kocev, Dejan Gjorgjevikj, and Saso Dzeroski. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084–3104, 2012.

[Matousek, 2008] Jirí Matousek. On variants of the Johnson-Lindenstrauss lemma. *Random Struct. Algorithms*, 33(2):142–156, 2008.

[Mendelson *et al.*, 2007] Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17(4):1248–1282, 2007.

[Prabhu and Varma, 2014] Yashoteja Prabhu and Manik Varma. FastXML: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *SIGKDD*, pages 263–272, August 2014.

[Rockafellar and Wets, 2004] R. Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*. Springer-Verlag Berlin Heidelberg, 2004.

[Shalev-Shwartz and Ben-David, 2014] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, 2014.

[Shawe-Taylor *et al.*, 1998] John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.

[Taha and Tiun, 2016] Adil Yaseen Taha and Sabrina Tiun. Binary relevance (BR) method classifier of multi-label classification for Arabic text. *Journal of Theoretical and Applied Information Technology*, 84(3):414–422, 2016.

[Tsoumakas *et al.*, 2010] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis P. Vlahavas. Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*, pages 667–685. 2010.

[Weinberger and Saul, 2009] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

[Yen *et al.*, 2016] Ian En-Hsu Yen, Xiangru Huang, Pradeep Ravikumar, Kai Zhong, and Inderjit S. Dhillon. PD-Sparse : A primal and dual sparse approach to extreme multiclass and multilabel classification. In *ICML*, pages 3069–3077, 2016.

[Yu *et al.*, 2014] Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit S. Dhillon. Large-scale multi-label learning with missing labels. In *ICML*, pages 593–601, 2014.

[Zhang and Schneider, 2012] Yi Zhang and Jeff G. Schneider. Maximum margin output coding. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1575–1582, 2012.

[Zhang and Zhou, 2007] Min-Ling Zhang and Zhi-Hua Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.