JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 27, Number 0, 2020

Pp. 1–18
DOI: 10.1089/cmb.2019.0329

Research Article

Learning Robust Multilabel Sample Specific Distances for Identifying HIV-1 Drug Resistance

LODEWIJK BRAND,[1] XUE YANG,[1] KAI LIU,[1] SAAD ELBELEIDY,[1] HUA WANG,[1] HAO ZHANG,[1] and FEIPING NIE[2]

ABSTRACT

AIDS is a syndrome caused by the HIV. During the progression of AIDS, a patient's immune system is weakened, which increases the patient's susceptibility to infections and diseases. Although antiretroviral drugs can effectively suppress HIV, the virus mutates very quickly and can become resistant to treatment. In addition, the virus can also become resistant to other treatments not currently being used through mutations, which is known in the clinical research community as cross-resistance. Since a single HIV strain can be resistant to multiple drugs, this problem is naturally represented as a multilabel classification problem. Given this multilabel relationship, traditional single-label classification methods often fail to effectively identify the drug resistances that may develop after a particular virus mutation. In this work, we propose a novel multilabel Robust Sample Specific Distance (RSSD) method to identify multiclass HIV drug resistance. Our method is novel in that it can illustrate the relative strength of the drug resistance of a reverse transcriptase (RT) sequence against a given drug nucleoside analog and learn the distance metrics for all the drug resistances. To learn the proposed RSSDs, we formulate a learning objective that maximizes the ratio of the summations of a number of ℓ_1-norm distances, which is difficult to solve in general. To solve this optimization problem, we derive an efficient, nongreedy iterative algorithm with rigorously proved convergence. Our new method has been verified on a public HIV type 1 drug resistance data set with over 600 RT sequences and five nucleoside analogs. We compared our method against several state-of-the-art multilabel classification methods, and the experimental results have demonstrated the effectiveness of our proposed method.

Keywords: drug resistance, HIV type 1, multilabel classification.

1. INTRODUCTION

According to estimations by the World Health Organization, around 35 million people are suffering from the HIV. HIV is a serious virus that attacks cells in the human immune system. During the later stages of the virus it can critically weaken the immune system and increase the patient's susceptibility to serious infection and disease. Fortunately, with the advent of antiretroviral therapies, we have been able to

[1]Department of Computer Science, Colorado School of Mines, Golden, Colorado.
[2]School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, P.R. China.

stem the progression of HIV and extend the life span of individuals affected by the virus. Unfortunately, the high mutation rates of HIV type 1 (HIV-1) can produce viral strains that adapt very quickly to new drugs (Smyth et al., 2012). The mutation of HIV-1 during antiretroviral treatments can lead to a phenomenon called ''cross-resistance'' (Heider et al., 2013; Riemenschneider et al., 2016). Cross-resistance of HIV-1 occurs when the virus develops resistance against the drugs, which are currently being used in addition to other drugs that have not yet been used in the treatment of a particular patient. This can make treatment of HIV-1 significantly more difficult, because a collection of drugs may not be effective after the initial treatment regimen due to the cross-resistance phenomenon observed in HIV-1. To address this problem, it is important that we develop automatic methods that can associate genetic strains of HIV to their corresponding drug resistances. The success of this research has the potential to reduce health care costs and increase the quality of life of those suffering from HIV and AIDS.

Recently, experimental testing of viral resistance in patients has been widely used in research, as well as in clinical settings to gain insight into the ways in which the drug resistance evolves. For example, large-scale pharmacogenomic screens have been conducted to explore the relationships between drug resistances and genomic sequences (Rhee et al., 2003). Besides, many clinical trials have been performed to discover mutation rates of the genetic subtypes of HIV-1 and how they develop resistances against various drug treatments (Pennings, 2012). In addition to these experimental phenotypic studies, computational approaches that use various machine learning methods offer the possibility to predict drug resistance in HIV-1 using short sequence information of the viral genotype, such as the genetic sequence of the viral reverse transcriptase (RT). For example, Rhee et al. (2006) used five different machine learning methods, including decision trees, artificial neural networks, support-vector machines, least-square regression, and least-angle regression, to investigate drug resistance in HIV-1 based on the RT sequences. Besides, genotype and phenotype features of HIV-1 extracted from RT sequences have been studied to predict drug resistance (Hepler et al., 2014).

In addition, a Bayesian algorithm (Gönen and Margolin, 2014) that combines kernel-based nonlinear dimensionality reduction and binary classification has been proposed to predict drug susceptibility of HIV within a multitask learning (MTL) framework. A critical drawback of these existing studies lies in the fact that they routinely consider HIV-1 drug resistance prediction as a *single-label* classification problem. This approach has been recognized to be inappropriate since HIV strains can develop resistances against multiple drugs at once due to their high mutation rate (Heider et al., 2013; Riemenschneider et al., 2016). To tackle this difficulty, following Heider et al., 2013 in this article, we solve the problem of HIV-1 drug resistance prediction as a *multilabel classification* problem.

Multilabel classification is an emerging research topic in machine learning driven by the advances of modern technologies in recent years (Wang et al., 2009, 2010a–c, 2015). As a generalization of traditional single-label classification that requires every data sample to belong to one and only one class, multilabel classification relaxes this restriction and allows one data sample to belong to multiple different classes at the same time. As a result, the classes in single-label classification problems are mutually exclusive, while those in multilabel classification problems are interdependent on one another. Although the labeling relaxation in multilabel classification problems have brought a number of successes in a variety of real-world applications (Wang et al., 2009, 2010c, 2015), it also causes labeling ambiguity that inevitably complicates the problem (Wang et al., 2010a,b).

In the context of predicting drug resistance developed by HIV-1, some HIV strains can develop the capability to resist multiple drugs, including those currently being used and those that have not yet been applied in a clinical setting. As a result, it is often unclear how to utilize a data sample that belongs to multiple classes to train a classifier for a given class (Wang et al., 2010a,b). A simple strategy to solve this problem is to use such data samples as the training data for all the classes to which they belong (Wang et al., 2009, 2010a), which is equivalent to assume that every data sample contributes equally to all their belonging classes when we train multilabel classification model (Wang et al., 2010b). However, this is not always the case in many real-world multilabel classification problems, which is particularly true in the problem of predicting drug resistance for HIV-1 because different mutations have different impact on resistance. Therefore, to create an effective multilabel classifier to predict HIV-1 resistances, it is critical to clarify the labeling ambiguity on data samples that belong to multiple classes and learn an appropriate scaling factor when we train the classifiers for different classes (Wang et al., 2010b).

In this study, we propose a novel Robust Sample Specific Distance (RSSD) for multilabel data to predict HIV-1 drug resistance, which, as illustrated in Figure 1, is able to explicitly rank the relevance of a training sample with respect to a specific class and characterize the second-order data-dependent statistics of all the

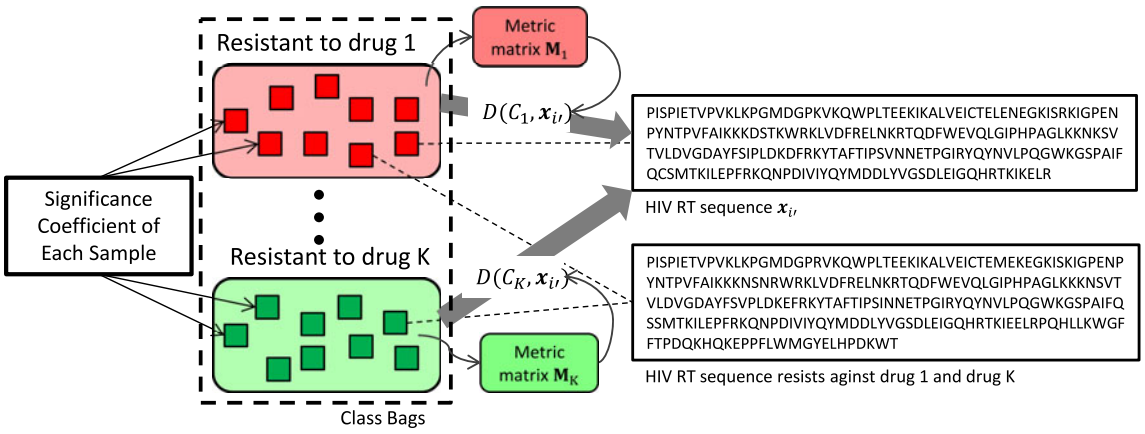

FIG. 1. The illustration of the proposed RSSD method. The small squares in the same color represent the data samples (RT sequences) that belong to one same class (e.g., resistance to a specific nucleoside analog). Two HIV RT sequences are listed in the right panel, which correspond to the data samples shown by the small squares connected by the dash lines. The top sequence in the right column only resists against drug 1, while the bottom sequence resists against both drug 1 and drug *K*, that is, it is a multilabel data sample. Ideally, the learned SCs for each data sample should be different with respect to different classes. For example, the bottom RT sequence is associated with s_{i1} for class 1 and s_{iK} for class K, which could be different depending on how the resistances evolved. RSSD, Robust Sample Specific Distance; RT, reverse transcriptase.

classes by class-wise distance metrics. In this study, we note that the proposed RSSD in this article is an application of the instance specific distance (ISD; Wang et al., 2011a–c, 2012, 2016) in single-instance (multilabel) classifications to solve the problem of predicting HIV-1 drug resistance, which was originally proposed in our previous works to solve multi-instance learning problems. We refer the interested readers to Wang et al. (2011a–c, 2012, 2016) for the definition of single-instance learning problems and that of multi-instance learning problems. To learn the sample relevances and the class-specific distance metrics, we formulate a learning objective that simultaneously maximizes and minimizes the summations of the ℓ_1-norm distances. To solve the optimization problem of our objective, using the same method in our recent works (Han et al., 2018; Liu et al., 2018), we derive an efficient iterative algorithm with theoretically guaranteed convergence, which, different from our previous works (Wang et al., 2012, 2014), is a *nongreedy* algorithm such that it has a better chance to find the optima of the proposed objective. We have applied our new method to predict the HIV-1 drug resistance on a public benchmark data set, and the experimental results have shown that our new RSSD method outperforms other state-of-the-art competing methods.

2. LEARNING RSSDS FOR MULTILABEL CLASSIFICATION

2.1. Notations and problem formalization

Throughout this article, we write matrices as bold uppercase letters and vectors as bold lowercase letters. The ℓ_1-norm of a vector $\mathbf{v}$ is defined as $\|\mathbf{v}\|_1 = \sum_i |v_i|$, and the ℓ_2-norm of $\mathbf{v}$ is defined as $\|\mathbf{v}\|_2 = \sqrt{\sum_i v_i^2}$. Given a matrix $\mathbf{M} = [m_{ij}]$, we denote its Frobenius norm as $\|\mathbf{M}\|_F$, and we define its ℓ_1-norm as $\|\mathbf{M}\|_1 = \sum_i \sum_j |m_{ij}|$. The trace of $\mathbf{M} = [m_{ij}]$ is defined as $\mathbf{tr}(\mathbf{M}) = \sum_i m_{ii}$.

In a multilabel classification problem, we are given a data set with n samples (n RT sequences) $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ and K classes (resistances to K target nucleoside analogs), where $\mathbf{x}_i \in \Re^d$ and $\mathbf{y}_i \in \{0, 1\}^K$, such that $\mathbf{y}_i(k) = 1$ if $\mathbf{x}_i$ belongs to the k-th class and $\mathbf{y}_i(k) = 0$ otherwise. Our goal is to learn from the training data $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ a classifier that is able to predict which nucleoside analogs (drug variants) a HIV-1 RT sequence is resistant to.

2.2. The class-to-sample distance

To learn the distance from a class to a data sample, we first represent each class as a bag consisting of all samples that belong to this class, that is, $C_k = \{\mathbf{x}_i | i \in \pi_k\}$, where π_k is the set of indices of all training samples that belong to the k-th class.

We first define the elementary distance from a data sample $\mathbf{x}_i$ in the k-th class bag C_k to another data sample $\mathbf{x}_{i'}$ as the squared Euclidean distance between the two involved vectors in the d-dimensional Euclidean space:

$$d_k(\mathbf{x}_i, \mathbf{x}_{i'}) = \|\mathbf{x}_i - \mathbf{x}_{i'}\|_2^2, \quad \forall\, i \in \pi_k,\ \forall\, k\ \ 1 \le k \le K. \tag{1}$$

We then compute the class-to-sample (C2S) distance from C_k to $\mathbf{x}_{i'}$ by summing all the elementary distances from the samples that belong to the k-th class to the data sample $\mathbf{x}_{i'}$:

$$D(C_k, \mathbf{x}_{i'}) = \sum_{\mathbf{x}_i \in C_k} d_k(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \mathbf{x}_{i'}\|_2^2. \tag{2}$$

2.3. Parameterized C2S distance

Because the C2S distance in Equation (2) does not take into account the resistance strength against a certain nucleoside analog, we further develop it by weighting the samples in a class bag by their relevance to this class.

Due to the ambiguous associations between the samples and the labels under the multilabel classification setting (Wang et al., 2010a,b), some samples in a class may characterize that particular class more strongly than the others from the statistical point of view. For example, Riemenschneider et al. (2016), where some viral RT sequences may develop a stronger drug resistance, while other viral RT sequences may be less resistant to a drug but may still be considered to be resistant. To develop an effective predictive model for HIV-1 drug resistance development, we need to capture these resistance differences. To be more specific, we should assign less weight to less resistant RT sequences when we determine whether or not to apply the ''resistant'' label to a query viral RT sequence.

Because we assume that counter-resistance against a target nucleoside analog does not exist, we define $s_{ik} \ge 0$ as a non-negative constant that assesses relative importance of $\mathbf{x}_i$ with respect to the k-th class, by which we can further develop the C2S distance as follows:

$$D(C_k, \mathbf{x}_{i'}) = \sum_{\mathbf{x}_i \in C_k} s_{ik}^2 \|\mathbf{x}_i - \mathbf{x}_{i'}\|_2^2. \tag{3}$$

Because s_{ik} reflects the relative importance of a sample $\mathbf{x}_i$ when we train a classifier for the k-th class, we call it the significance coefficient (SC) of $\mathbf{x}_i$ with respect to the k-th class. Obviously, the SCs quantitatively assess the resistances developed by the training of viral RT sequences against the target nucleoside analogs during the learning process.

2.4. Parameterized C2S distance refined by class specific distance metrics

The RSSD defined in Equation (3) is simply a weighted Euclidean distance that does not take into account the information conveyed by the input data other than the first-order statistics. Similar to many other statistical models in machine learning, using the Mahalanobis distances with appropriate distance metrics is recommended to capture the second-order statistics of the input data. Instead of learning one single global distance metric for all the classes as in many existing statistical studies, we propose to learn K different class-specific distance metrics $\{\mathbf{M}_k \succ 0\}_{k=1}^K \in \Re^{d\times d}$, one for each class. Thus we further develop the parameterized C2S distance as:

$$D(C_k, \mathbf{x}_{i'}) = \sum_{\mathbf{x}_i \in C_k} s_{ik}^2 (\mathbf{x}_i - \mathbf{x}_{i'})^T \mathbf{M}_k \left(\mathbf{x}_i - \mathbf{x}'_i\right). \tag{4}$$

Because the class-specific distance metric $\mathbf{M}_k$ is a positive definite matrix, that is, $\mathbf{M} \succ 0$, we can reasonably write it as $\mathbf{M}_k = \mathbf{W}_k \mathbf{W}_k^T$, where $\mathbf{W}_k \in \Re^{d\times r}$ is an orthonormal matrix such that $\mathbf{W}_k^T \mathbf{W}_k = \mathbf{I}$. Thus we can rewrite Equation (4) as follows:

$$D(C_k, \mathbf{x}_{i'}) = \sum_{\mathbf{x}_i \in C_k} s_{ik}^2 (\mathbf{x}_i - \mathbf{x}_{i'})^T \mathbf{W}_k \mathbf{W}_k^T (\mathbf{x}_i - \mathbf{x}_{i'}) = \sum_{\mathbf{x}_i \in C_k} \left\|\mathbf{W}_k^T (\mathbf{x}_i - \mathbf{x}_{i'}) s_{ik}\right\|_2^2. \tag{5}$$

A critical problem of $D(C_k, \mathbf{x}_{i'})$ defined in Equation (5) lies in that it computes the summation of a number of squared ℓ_2-norm distances. These squared terms are notoriously known to be sensitive to both

outlying samples and features (Ding et al., 2006; Wang et al., 2014; Liu et al., 2019a,b; Yang et al., 2019). Due to the cross-resistance phenomenon (Heider et al., 2013), this problem is particularly significant for identifying HIV-1 drug resistance. To promote the robustness of $D(C_k, \mathbf{x}_{i'})$ against outliers, following many previous works (Ke and Kanade, 2005; Ding et al., 2006; Kwak, 2008; Wright et al., 2009; Wang et al., 2013b, 2014; Liu et al., 2019a,b; Yang et al., 2019), we define it using the ℓ_1-norm distance as follows:

$$D(C_k, \mathbf{x}_{i'}) = \sum_{\mathbf{x}_i \in C_k} \left\| \mathbf{W}_k^T (\mathbf{x}_i - \mathbf{x}_{i'}) s_{ik} \right\|_1, \tag{6}$$

which we call the proposed *RSSD*.

To use RSSD defined in Equation (6), we need to learn two sets of parameters s_{ik} and $\mathbf{W}_k$ for every class, where we use the learned $\mathbf{W}_k$ to compute the metric matrix as $\mathbf{M}_k = \mathbf{W}_k \mathbf{W}_k^T$. Following the most broadly used machine learning strategy to maximize data discriminativity for classification, such as Fisher's linear discriminant (Fukunaga, 2013), for a given class C_k we simultaneously maximize the overall RSSDs from every class bag to all its nonbelonging samples and minimize the overall RSSDs from every class bag to all the samples belonging to that class:

$$\max_{\mathbf{W}_k,\, s_{ik}} \frac{\sum_{\mathbf{x}_{i'} \notin C_k} \sum_{\mathbf{x}_i \in C_k} \left\| \mathbf{W}_k^T (\mathbf{x}_i - \mathbf{x}_{i'}) s_{ik} \right\|_1}{\sum_{\mathbf{x}_{i'} \in C_k} \sum_{\mathbf{x}_i \in C_k} \left\| \mathbf{W}_k^T (\mathbf{x}_i - \mathbf{x}_{i'}) s_{ik} \right\|_1}, \quad s.t.\ \mathbf{W}_k^T \mathbf{W}_k = \mathbf{I},\ s_{ik} \geq 0. \tag{7}$$

Learning the RSSDs by solving Equation (7) and classifying query viral RT sequences using the adaptive decision boundary method (Wang et al., 2009, 2013a), our proposed RSSD method can be used for multilabel classification.

3. AN EFFICIENT SOLUTION ALGORITHM

Our new objective in Equation (7) maximizes the ratio of the summations of a number of ℓ_1-norm distances, which is obviously not smooth thereby making it difficult to solve in general. To solve this challenging optimization problem, we first turn to solve the following generalized objective:

$$v_{\text{opt}} = \arg\max_{v \in \Omega} \frac{h(v)}{m(v)}, \qquad \forall v \in \boldsymbol{\Omega} \quad \begin{cases} C_2 \geq m(v) \geq C_1 > 0, \\ C_4 \geq h(v) \geq C_3 > 0. \end{cases} \tag{8}$$

where Ω is the feasible domain. Next, we propose a simple, yet efficient, iterative framework in Algorithm 1 to solve the objective in Equation (8). The convergence of Algorithm 1 is rigorously guaranteed by Theorem 1.

Algorithm 1: Algorithm to solve Equation (8).

1. Randomly initialize $v^0 \in \Omega$ and set $t = 1$.
while *not converge* **do**
  2. Calculate $\lambda^t = \frac{h(v^{t-1})}{m(v^{t-1})}$.
  3. Find a $v^t \in \Omega$ satisfying $h(v^t) - \lambda^t m(v^t) > h(v^{t-1}) - \lambda^t m(v^{t-1}) = 0$.
  4. $t = t + 1$.
Output: v.

Theorem 1. *In Algorithm 1, for each iteration we have* $\frac{h(v^t)}{m(v^t)} \geq \frac{h(v^{t-1})}{m(v^{t-1})}$ *and* $\forall \delta$*, there must exist a* $\hat{t}$ *such that* $\forall t > \hat{t}\ \frac{h(v^t)}{m(v^t)} - \frac{h(v^{t-1})}{m(v^{t-1})} < \delta$.

Proof. In Algorithm 1, from step 3 we have $h(v^t) - \lambda^t m(v^t) > 0$. Because $\forall v \in \boldsymbol{\Omega}\ \ m(v) > 0$, we can get $\frac{h(v^t)}{m(v^t)} > \lambda^t = \frac{h(v^{t-1})}{m(v^{t-1})}$, which completes the proof of the first statement of Theorem 1.

Suppose that for the k-th iteration, there exists a c^t such that $h(v^t) - \lambda^t m(v^t) = c^t > 0$. We have:

$$\frac{h(v^t)}{m(v^t)} = \frac{h(v^{t-1})}{m(v^{t-1})} + \frac{c_t}{m(v^t)}, \tag{9}$$

by which we can derive:

$$\frac{h(v^t)}{m(v^t)} = \frac{h(v^0)}{m(v^0)} + \sum_{i=1}^{t} \frac{c^i}{m(v^i)}. \tag{10}$$

From Equation (10), we can derive:

$$\frac{h(v^0)}{m(v^0)} + \frac{1}{C_2}\sum_{i=1}^{t} c^i \leq \frac{h(v^t)}{m(v^t)} \leq \frac{h(v^0)}{m(v^0)} + \frac{1}{C_1}\sum_{i=1}^{t} c^i. \tag{11}$$

Suppose that there exist a positive constant C such that $\lim_{t\to\infty}\sum_{i=1}^{t} c^i = C$. If this is not true, we have $\lim_{t\to\infty}\sum_{i=1}^{t} c^i = \infty$, by which, together with Equation (11), we can derive $\lim_{t\to\infty}\sum_{i=1}^{t}\frac{h(v^t)}{m(v^t)} = \infty$. This, however, contradicts the fact that $\frac{h(v^t)}{m(v^t)}$ is bounded as defined in Equation (8), which means that the following holds:

$$\lim_{t\to\infty}\sum_{i=1}^{t} c^i = C. \tag{12}$$

Thus, we have:

$$\lim_{t\to\infty} c^t = 0, \tag{13}$$

which means that:

$$\lim_{t\to\infty}\frac{c^t}{m(v^t)} = 0, \tag{14}$$

which indicates that: $\forall\delta > 0$, there must exist a $\hat{t}$ such that:

$$\forall t > \hat{t} \quad \frac{c^t}{m(v^t)} < \delta, \tag{15}$$

by which and Equation (9), we have:

$$\forall t > \hat{t} \quad \frac{h(v^t)}{m(v^t)} - \frac{h(v^{t-1})}{m(v^{t-1})} < \delta, \tag{16}$$

which indicates that Algorithm 1 converges to a local optimum and completes the proof of the second statement of Theorem 1. ■

3.1. *Fixing* s_{ik} *to solve* $\mathbf{W}_k$

According to step 3 in Algorithm 1, we can easily write the corresponding inequality of our objective in Equation (7) as:

$$F(\mathbf{W}_k) = H(\mathbf{W}_k) - \lambda^t M(\mathbf{W}_k) \geq 0, \tag{17}$$

where λ^t is computed by

$$\lambda^t = \frac{\sum_{\mathbf{x}_i'\notin C_k}\sum_{\mathbf{x}_i\in C_k}\left\|(\mathbf{W}_k^{t-1})^T(\mathbf{x}_i - \mathbf{x}_{i'})s_{ik}\right\|_1}{\sum_{\mathbf{x}_i'\in C_k}\sum_{\mathbf{x}_i\in C_k}\left\|(\mathbf{W}_k^{t-1})^T(\mathbf{x}_i - \mathbf{x}_{i'})s_{ik}\right\|_1}, \tag{18}$$

and

$$H(\mathbf{W}_k) = \sum_{\mathbf{x}_i'\notin C_k}\sum_{\mathbf{x}_i\in C_k}\left\|\mathbf{W}_k^T(\mathbf{x}_i - \mathbf{x}_{i'})s_{ik}\right\|_1, \tag{19}$$

$$M(\mathbf{W}_k) = \sum_{\mathbf{x}_i'\in C_k}\sum_{\mathbf{x}_i\in C_k}\left\|\mathbf{W}_k^T(\mathbf{x}_i - \mathbf{x}_{i'})s_{ik}\right\|_1. \tag{20}$$

In Equation (18), $\mathbf{W}_k^{t-1}$ denotes the projection matrix in the $(t-1)$-th iteration.

Now we need to solve the problem in Equation (17), for which we first introduce the following two lemmas:

Lemma 1. (Liu et al., 2017, theorem 1) *For any vector* $\boldsymbol{\xi}=[\xi_1, \cdots, \xi_m]^T \in \Re^m$, *we have* $\|\boldsymbol{\xi}\|_1 = \max_{\boldsymbol{\eta}\in\Re^m} (\text{sign}(\boldsymbol{\eta}))^T \boldsymbol{\xi}$, *where the maximum value is attained if and only if* $\boldsymbol{\eta}=a\times\boldsymbol{\xi}$, *where* $a>0$ *is a scalar.*

Lemma 2. (Jenatton et al., 2010, lemma 3.1) *For any vector* $\boldsymbol{\xi}=[\xi_1, \cdots, \xi_m]^T \in \Re^m$, *we have* $\|\boldsymbol{\xi}\|_1 = \min_{\boldsymbol{\eta}\in\Re^m_+} \frac{1}{2}\sum_{i=1}^{m}\frac{\xi_i^2}{\eta_i} + \frac{1}{2}\|\boldsymbol{\eta}\|_1$, *where the minimum value is attained if and only if* $\eta_j = |\xi_j|, j \in \{1, 2, \cdots, m\}$.

Motivated by Lemmas 1 and 2, we construct the following objective:

$$L(\mathbf{W}_k, \mathbf{W}_k^{t-1}) = K(\mathbf{W}_k) - \lambda^t N(\mathbf{W}_k), \tag{21}$$

where $K(\mathbf{W}_k)$ and $N(\mathbf{W}_k)$ are defined as:

$$K(\mathbf{W}_k) = \sum_{g=1}^{r} \mathbf{w}_g^T \mathbf{B}\text{sign}\left(\mathbf{B}^T \mathbf{w}_g^{t-1}\right), \tag{22}$$

$$N(\mathbf{W}_k) = \frac{1}{2}\sum_{g=1}^{r} \mathbf{w}_g^T \mathbf{A}_g \mathbf{w}_g + \left(\mathbf{w}_g^{t-1}\right)^T \mathbf{A}_g \mathbf{w}_g^{t-1}. \tag{23}$$

Here $\mathbf{w}_g$ and $\mathbf{w}_g^{t-1}$ denote the g-th column of matrices $\mathbf{W}_k$ and $\mathbf{W}_k^{t-1}$, respectively; $\mathbf{B}$ and $\mathbf{A}_g$ for $g=1, 2, \cdots, r$ are defined as follows:

$$\mathbf{B} = [\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}, \overline{\mathbf{x}}_2 - \overline{\mathbf{x}}, \cdots, \overline{\mathbf{x}}_n - \overline{\mathbf{x}}], \tag{24}$$

$$\mathbf{A}_g = \sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \frac{\left(\mathbf{x}_j - \overline{\mathbf{x}}_i\right)\left(\mathbf{x}_j - \overline{\mathbf{x}}_i\right)^T}{\left|\left(\mathbf{w}_g^{t-1}\right)^T \left(\mathbf{x}_j - \overline{\mathbf{x}}_i\right)\right|}, \tag{25}$$

and sign(x) is the sign function, which is defined as follows:

$$\text{sign}(x) = \begin{cases} 1 & \text{if} x \geq 0 \\ -1 & \text{if} x < 0 \end{cases}. \tag{26}$$

Then, using the definition of $L(\mathbf{W}_k, \mathbf{W}_k^{t-1})$ in Equation (21) and Lemmas 1 to 2, we can prove the following theorem.

Theorem 2. *For any* $\mathbf{W}_k \in \Re^{d\times r}$, *we have:*

$$L(\mathbf{W}_k, \mathbf{W}_k^{t-1}) \leq F(\mathbf{W}_k), \tag{27}$$

where the equality holds if and only if $\mathbf{W}_k = \mathbf{W}_k^{t-1}$.

Proof. First, according to Lemma 1 we can compute:

$$\begin{aligned} & H(\mathbf{W}_k) \\ &= \sum_{i=1}^{n} \left\|\mathbf{W}_k^T(\mathbf{x}_i - \overline{\mathbf{x}})\right\|_1 \\ &= \sum_{i=1}^{n}\sum_{g=1}^{r} \left\|\mathbf{w}_g^T(\mathbf{x}_i - \overline{\mathbf{x}})\right\|_1 \\ &\geq \sum_{g=1}^{r}\sum_{i=1}^{n} \text{sign}\left[(\mathbf{w}_g^{k-1})^T(\mathbf{x}_i - \overline{\mathbf{x}})\right]\left[\mathbf{w}_g^T(\mathbf{x}_i - \overline{\mathbf{x}})\right] \\ &= \sum_{g=1}^{r} \mathbf{w}_g^T \mathbf{B}\text{sign}\left(\mathbf{B}^T \mathbf{w}_g^{k-1}\right) \\ &= K(\mathbf{W}_k). \end{aligned} \tag{28}$$

Then, according to Lemma 2 we have:

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\{\mathcal{N}_i\cup\{\mathbf{x}_i\}\}}\left\{\frac{1}{2}\frac{\boldsymbol{\xi}^T(\mathbf{x}_j-\overline{\mathbf{x}}_i)(\mathbf{x}_j-\overline{\mathbf{x}}_i)^T\boldsymbol{\xi}}{\boldsymbol{\xi}^T(\mathbf{x}_j-\overline{\mathbf{x}}_i)}+\frac{1}{2}\left\|\boldsymbol{\xi}^T(\mathbf{x}_j-\overline{\mathbf{x}}_i)\right\|_1\right\}\le\\&\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\{\mathcal{N}_i\cup\{\mathbf{x}_i\}\}}\left\{\frac{1}{2}\frac{\boldsymbol{\xi}^T(\mathbf{x}_j-\overline{\mathbf{x}}_i)(\mathbf{x}_j-\overline{\mathbf{x}}_i)^T\boldsymbol{\xi}}{\boldsymbol{\eta}^T(\mathbf{x}_j-\overline{\mathbf{x}}_i)}+\frac{1}{2}\left\|\boldsymbol{\eta}^T(\mathbf{x}_j-\overline{\mathbf{x}}_i)\right\|_1\right\},\end{aligned}\tag{29}$$

which indicates that:

$$\begin{aligned}&M(\mathbf{W}_k)\\&=\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\{\mathcal{N}_i\cup\{\mathbf{x}_i\}\}}\left\|\mathbf{W}_k^T(\mathbf{x}_j-\overline{\mathbf{x}}_i)\right\|_1\\&=\sum_{g=1}^{r}\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\{\mathcal{N}_i\cup\{\mathbf{x}_i\}\}}\left\{\frac{1}{2}\frac{\mathbf{w}_g^T(\mathbf{x}_j-\overline{\mathbf{x}}_i)(\mathbf{x}_j-\overline{\mathbf{x}}_i)^T\mathbf{w}_g}{\mathbf{w}_g^T(\mathbf{x}_j-\overline{\mathbf{x}}_i)}+\frac{1}{2}\left\|\mathbf{w}_g^T(\mathbf{x}_j-\overline{\mathbf{x}}_i)\right\|_1\right\}\\&\le\sum_{g=1}^{r}\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\{\mathcal{N}_i\cup\{\mathbf{x}_i\}\}}\left\{\frac{1}{2}\frac{\mathbf{w}_g^T(\mathbf{x}_j-\overline{\mathbf{x}}_i)(\mathbf{x}_j-\overline{\mathbf{x}}_i)^T\mathbf{w}_g}{\left(\mathbf{w}_g^{k-1}\right)^T(\mathbf{x}_j-\overline{\mathbf{x}}_i)}+\frac{1}{2}\left\|\left(\mathbf{w}_g^{k-1}\right)^T(\mathbf{x}_j-\overline{\mathbf{x}}_i)\right\|_1\right\}\\&=\frac{1}{2}\sum_{g=1}^{r}\mathbf{w}_g^T\mathbf{A}_g\mathbf{w}_g+\left(\mathbf{w}_g^{k-1}\right)^T\mathbf{A}_g\mathbf{w}_g^{k-1}\\&=N(\mathbf{W}_k).\end{aligned}\tag{30}$$

Combining Equations (28) and (30), we can derive:

$$L(\mathbf{W}_k,\mathbf{W}_k^{t-1})=K(\mathbf{W}_k)-\lambda^k N(\mathbf{W}_k)\le H(\mathbf{W}_k)-\lambda^k M(\mathbf{W}_k)=F(\mathbf{W}_k).\tag{31}$$

According to Lemmas 1 and 2, it is easy to verify that equality holds in Equations (28) and (30) if and only if $\mathbf{W}_k=\mathbf{W}_k^{t-1}$. Thus, equality holds in Equation (31) if and only if $\mathbf{W}_k=\mathbf{W}_k^{t-1}$. This completes the proof of Theorem 2. ■

Now we continue to solve our objective. Let $\mathbf{W}_k=\mathbf{W}_k^{t-1}$, by substituting it into the objective, we have $L(\mathbf{W}_k,\mathbf{W}_k^{k-1})=F(\mathbf{W}_k^{t-1})=0$. In the k-th iteration in solving the objective in Equation (7), $\mathbf{W}_k^\star$ satisfies $L(\mathbf{W}_k^\star,\mathbf{W}_k^{t-1})\ge L(\mathbf{W}_k^{t-1},\mathbf{W}_k^{t-1})=0$. Then, we have

$$F(\mathbf{W}_k^\star)\ge L(\mathbf{W}_k^\star,\mathbf{W}_k^{t-1})\ge L(\mathbf{W}_k^{t-1},\mathbf{W}_k^{t-1})=F(\mathbf{W}_k^{t-1})=0.\tag{32}$$

Lemma 1 and Equation (32) indicate that the solution of the objective function in Equation (17) can be transformed to solve the objective function $L(\mathbf{W}_k,\mathbf{W}_k^{t-1})\ge 0$, which can be easily solved by the projected subgradient method with Armijo line search (Sun and Yuan, 2006). Note that, for any matrix $\mathbf{W}_k$ the operator $P(\mathbf{W}_k)=\mathbf{W}_k\left(\mathbf{W}_k^T\mathbf{W}_k\right)^{-\frac{1}{2}}$ can project it onto an orthogonal cone. This guarantees the orthogonality constraint of the projection matrix, that is, $\left(\mathbf{W}_k^t\right)^T\left(\mathbf{W}_k^t\right)=\mathbf{I}$. Algorithm 2 summarizes the algorithm to solve the objective in Equation (17).

Algorithm 2: Algorithm to maximize $F(\mathbf{W}_k)$.

Input: $\mathbf{W}_k^{t-1}$ and Armijo parameter $0<\beta<1$.
1. Calculate λ^k by Equation (18).
2. Calculate the subgradient $\mathbf{G}^{k-1}=\partial L(\mathbf{W}_k^{t-1},\mathbf{W}_k^{t-1})=\mathbf{B}\ \mathrm{sign}\left(\mathbf{B}^T\mathbf{W}_k^{t-1}\right)-\lambda^k[\mathbf{A}_1\mathbf{w}_1,\mathbf{A}_2\mathbf{w}_2,\cdots,\mathbf{A}_r\mathbf{w}_r]$.
3. Set $t=1$.
while *not* $F(\mathbf{W}_k^t)>F(\mathbf{W}_k^{t-1})=0$ **do**
  4. Calculate $\mathbf{W}_k^t=P(\mathbf{W}_k^{t-1}+\beta^m\mathbf{G}^{t-1})$.
  5. Calculate $F(\mathbf{W}_k^t)$ by Equation (17).
  6. $t=t+1$.
Output: $\mathbf{W}_k^k$.

Finally, based on Algorithm 2, we can derive a simple yet efficient iterative algorithm as summarized in Algorithm 3 to solve our objective in Equation (7) when s_{ik} is fixed. In addition, Theorem 3 indicates that our proposed Algorithm 3 monotonically increases the objective function value in each iteration. Theorem 4 indicates that the objective function is upper bounded, which, together with Theorem 3, indicates that Algorithm 3 converges to a local optimum.

Algorithm 3: Algorithm for nongreedy ratio maximization of the ℓ_1-norm distances.

1. Randomly initialize $\mathbf{W}_k^0$ satisfying $\left(\mathbf{W}_k^0\right)^T\mathbf{W}_k^0=\mathbf{I}$ and set $t=1$.
while *not converge* **do**
2. Calculate λ^t by Equation (18).
3. Find a $\mathbf{W}_k^t$ satisfying $F(\mathbf{W}_k^t) > F(\mathbf{W}_k^{t-1})=0$ by Algorithm 2.
4. $t=t+1$.
Output: $\mathbf{W}$.

Theorem 3. *If $\mathbf{W}_k^t$ is the solution of the objective function in Equation (17) and satisfies $\left(\mathbf{W}_k^t\right)^T\left(\mathbf{W}_k^t\right)=\mathbf{I}$, then we have $\mathcal{J}(\mathbf{W}_k^t) \geq \mathcal{J}(\mathbf{W}_k^{t-1})$.*

Proof. Since $\mathbf{W}_k^k$ is the solution of the objective function in Equation (17), we have

$$F(\mathbf{W}_k^t)=\sum_{i=1}^{n}\left\|\left(\mathbf{W}_k^t\right)^T(\mathbf{x}_i-\overline{\mathbf{x}})\right\|_1-\lambda^k\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\{\mathcal{N}_i\cup\{\mathbf{x}_i\}\}}\left\|\left(\mathbf{W}_k^t\right)^T\left(\mathbf{x}_j-\overline{\mathbf{x}}_i\right)\right\|_1\geq 0, \tag{33}$$

from which we can easily derive:

$$\mathcal{J}(\mathbf{W}_k^t)=\frac{\sum_{i=1}^{n}\left\|\left(\mathbf{W}_k^t\right)^T(\mathbf{x}_i-\overline{\mathbf{x}})\right\|_1}{\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\{\mathcal{N}_i\cup\{\mathbf{x}_i\}\}}\left\|\left(\mathbf{W}_k^t\right)^T\left(\mathbf{x}_j-\overline{\mathbf{x}}_i\right)\right\|_1}\geq\lambda^k. \tag{34}$$

Now by substituting Equation (18) into Equation (34), we have

$$\begin{aligned}\mathcal{J}(\mathbf{W}_k^t)&=\frac{\sum_{i=1}^{n}\left\|\left(\mathbf{W}_k^t\right)^T(\mathbf{x}_i-\overline{\mathbf{x}})\right\|_1}{\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\{\mathcal{N}_i\cup\{\mathbf{x}_i\}\}}\left\|\left(\mathbf{W}_k^t\right)^T\left(\mathbf{x}_j-\overline{\mathbf{x}}_i\right)\right\|_1}\\&\geq\frac{\sum_{i=1}^{n}\left\|\left(\mathbf{W}_k^{t-1}\right)^T(\mathbf{x}_i-\overline{\mathbf{x}})\right\|_1}{\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\{\mathcal{N}_i\cup\{\mathbf{x}_i\}\}}\left\|\left(\mathbf{W}_k^{t-1}\right)^T\left(\mathbf{x}_j-\overline{\mathbf{x}}_i\right)\right\|_1}\end{aligned} \tag{35}$$

$$=\mathcal{J}(\mathbf{W}_k^{t-1}), \tag{36}$$

which completes the proof of Theorem 3. ■

Theorem 4. *The objective in Equation* (7) *is upper bounded.*

Proof. First, using Cauchy–Schwarz inequality we have the following for the numerator of our objective in Equation (7):

$$\begin{aligned}
&\sum_{i=1}^{n} \left\| \mathbf{W}_k^T (\mathbf{x}_i - \overline{\mathbf{x}}) \right\|_1 \\
&= \sum_{i=1}^{n} \sum_{j=1}^{r} \left\| \mathbf{w}_j^T (\mathbf{x}_i - \overline{\mathbf{x}}) \right\|_1 \\
&\leq \sum_{i=1}^{n} \sum_{j=1}^{r} \left\| \mathbf{w}_j^T \right\|_2 \| (\mathbf{x}_i - \overline{\mathbf{x}}) \|_2 \\
&= \sum_{i=1}^{n} r \| (\mathbf{x}_i - \overline{\mathbf{x}}) \|_2 .
\end{aligned} \tag{37}$$

Obviously, given an input data set, $\sum_{i=1}^{n} r \| (\mathbf{x}_i - \overline{\mathbf{x}}) \|_2$ is a constant, which indicates that the numerator of our objective in Equation (7) is upper bounded for a given data set.

Second, it can be verified that $\sqrt{\sum_{i=1}^{n} v_i^2} \leq \sum_{i=1}^{n} |v_i|$, that is, $\forall \mathbf{v} \in R^n \ \|\mathbf{v}\|_2 \leq \|\mathbf{v}\|_1$, by which we can derive the following for the denominator of our objective in Equation (7):

$$\begin{aligned}
&\sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \left\| \mathbf{W}_k^T (\mathbf{x}_j - \overline{\mathbf{x}}_i) \right\|_1 \\
&\geq \sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \sqrt{\left\| \mathbf{W}_k^T (\mathbf{x}_j - \overline{\mathbf{x}}_i) \right\|_2^2} \\
&\geq \sqrt{\sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \left\| \mathbf{W}_k^T (\mathbf{x}_j - \overline{\mathbf{x}}_i) \right\|_2^2} \\
&= \sqrt{\mathbf{tr}\left(\mathbf{W}_k^T \mathbf{S}_L \mathbf{W}_k \right)} \\
&\geq \sqrt{\sum_{i=1}^{r} \lambda_i},
\end{aligned} \tag{38}$$

where $\lambda_i (i = 1, \ldots, r)$, ordered by $\lambda_1 \leq \cdots \leq \lambda_r$, are the eigenvalues of $\mathbf{S}_L$. The last inequality in Equation (38) is obtained by the Ky Fan's inequality (Fan, 1950), which states that $\mathbf{tr}\left(\mathbf{W}_k^T \mathbf{S}_L \mathbf{W}_k \right) \geq \sum_{i=1}^{r} \lambda_i$. Again, given an input data set, $\mathbf{S}_L$ is a constant matrix thereby $\sum_{i=1}^{r} \lambda_i$ is a constant. Thus the denominator of our objective in Equation (7) is lower bounded.

The two bounds in Equations (37) and (38) together indicate that our objective in Equation (7) is upper bounded. ■

3.2. *Fixing* $\mathbf{W}_k$ *to solve* $\mathrm{s_{ik}}$

When $\mathbf{W}_k$ is fixed, we define a scalar $d_{ii'k} = \left\| \mathbf{W}_k^T (\mathbf{x}_i - \mathbf{x}_{i'}) \right\|_1$. Then we write Equation (7) as:

$$\max \frac{\sum_{\mathbf{x}_i' \notin C_k} \sum_{\mathbf{x}_i \in C_k} s_{ik} d_{ii'k}}{\sum_{\mathbf{x}_i' \in C_k} \sum_{\mathbf{x}_i \in C_k} s_{ik} d_{ii'k}}, \quad s.t.\ s_{ik} \geq 0. \tag{39}$$

Defining that $d_{ik}^w = \sum_{i' \in \pi_k} d_{ii'k}$ and $d_{ik}^b = \sum_{i' \notin \pi_k} d_{ii'k}$, we can further rewrite the objective as:

$$\max \frac{\sum_{\mathbf{x}_i \notin C_k} s_{ik} d_{ik}^w}{\sum_{\mathbf{x}_i \in C_k} s_{ik} d_{ik}^b}, \quad s.t.\ s_{ik} \geq 0. \tag{40}$$

Again, to solve Equation (40), to step 3 in Algorithm 1, we solve the following optimization problem:

$$\max \sum_{\mathbf{x}_i \in C_k} s_{ik} d_{ik}^w - \lambda \sum_{\mathbf{x}_i \in C_k} s_{ik} d_{ik}^b, \quad s.t.\ s_{ik} \geq 0, \tag{41}$$

where λ is computed as Equation (18) in the t-th iteration.

Define that $d_{ik} = d_{ik}^w - \lambda d_{ik}^b$, we can rewrite the optimization problem in Equation (41) as:

$$\max \sum_{\mathbf{x}_i \in C_k} s_{ik} d_{ik}, \quad s.t.\ s_{ik} \geq 0, \tag{42}$$

The problem in Equation (42) can be decoupled to solve the following subproblems separately for each $\mathbf{x}_i \in C_k$:

$$\max\ s_{ik} d_{ik}, \quad s.t.\ s_{ik} \geq 0, \tag{43}$$

which is a convex linear programming problem (Wright and Nocedal, 1999) and can be solved efficiently by many off-the-shelf solution algorithms (Wright and Nocedal, 1999). By inserting the solution to Equation (43) after step 3 of Algorithm 3, we can finally solve our objective in Equation (7), which is equivalent to performing alternative optimization. Therefore, the algorithm is guaranteed to converge to a local optimum.

4. EXPERIMENTAL RESULTS

We evaluate the proposed RSSD method using a publicly available HIV drug resistance database (Rhee et al., 2006), which contains HIV-1 RT sequences with associated resistance factors measured by IC_{50} ratios. We analyze the drug resistance of these RT sequences against five nucleoside analogs: Lamivudine (3TC), Abacavir (*ABC*), Zidovudine (*AZT*), Stavudine (*d4T*), and Didanosine (*ddI*). Following Heider et al. (2010), although the Tenofovir nucleoside analog is included in this database, it is not used in our study, because the number of the RT sequences resistant to this nucleoside analog is very small. As a result, we end up with 623 RT sequences for our experiments.

Drug resistance of a particular HIV strain is measured by the IC_{50} ratio, which is defined as the concentration of a specific drug inhibiting 50% of viral replication compared with cell culture experiments without the drug:

$$\frac{IC_{50}(\text{drug concentration for resistant strain})}{IC_{50}(\text{drug concentration for wild type})}. \tag{44}$$

We label the viral RT sequences as ''resistant'' using the same drug-specific IC_{50} ratio cutoff thresholds as in Heider et al. (2013), which are set to 3.0 for 3TC and *AZT*, 2.0 for *ABC*, and 1.5 for *ddI* and *d4T*. We use hydrophobicity characteristics (Kyte and Doolittle, 1982) to represent the RT sequences, which have demonstrated good prediction performance in many protein classification studies (Heider et al., 2010). For each RT sequence, we extract a hydrophobicity vector, which is obtained from the amino acid sequence and smoothed within a window. The length of the original hydrophobicity vectors may be different due to the different lengths of the RT sequences. In this study, following Heider et al. (2013) we set a fixed window size of 11 and interpolated all hydrophobicity vectors to length 230 using the spline interpolation method (Kyte and Doolittle, 1982).

4.1. Convergence of the proposed algorithm

In Section 3, we have theoretically proved the convergence of the derived solution algorithm. Now we study the convergence of our new algorithm from the empirical perspective. We apply our new method on the HIV-1 drug resistance data set and plot the objective value after each iteration in Figure 2. This plot clearly shows that our solution algorithm converges very fast and confirms the correctness of our new method.

4.2. Parameter selection of the proposed method

Predicting drug resistance for HIV-1 RT sequences is a multilabel classification problem. Therefore, we evaluate the proposed method by two broadly used multilabel performance metrics (Lewis et al., 2004): Hamming loss and average precision. The Hamming loss is computed over all instances over all classes. The average precision is calculated for both the micro and macro averages. In multilabel classification, the

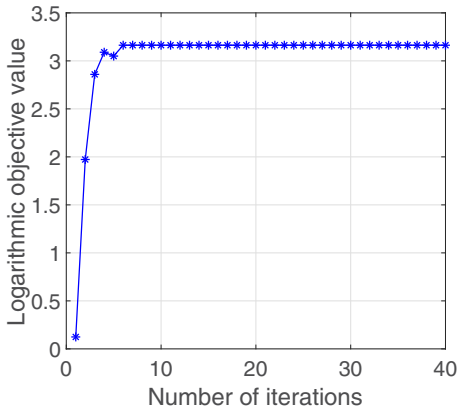

FIG. 2. The convergence of our solution algorithm.

macro average is computed as the average of the precision values over all the classes; thus, it attributes equal weights to every individual class. In contrast, the micro average is obtained from the summation of contingency matrices for all binary classifiers, thus it gives equal weight to all classifiers and emphasizes the accuracy of categories with more positive samples.

The proposed RSSD has only one parameter: the dimensionality r of $\mathbf{W}_k$. Ideally, each class can have its own fine tuned parameter. To reduce the experimental effort, we fix the parameter r across all classes in our studies. We first evaluate the impacts of the parameter in a standard fivefold cross-validation experiment, where we select r in the range from 10 to 100. The classification performance measured by the three aforementioned multilabel performance metrics, when we vary r, is reported in Figure 3. The results in these experiments show that the classification performance of the proposed method is reasonably stable when we vary r in a considerably large selection range. This illustrates that tuning parameters in our proposed method is not a difficult task, which adds to the practical value of our method to solve real-world problems. Based on these observations, we fix $r=50$ in all our future experiments for simplicity.

4.3. Comparative studies

We use a standard fivefold cross-validation to evaluate the predictive capability of the proposed RSSD method. We implement two versions of our proposed method, one version that defines $D(C_k, \mathbf{x}_{i'})$ using the ℓ_1-norm distances as in Equation (6; denoted as ''Ours-ℓ_1'') and another that defines $D(C_k, \mathbf{x}_{i'})$ using the squared ℓ_2-norm distances as in Equation (5; denoted as ''Ours-ℓ_2^2''). We compare our new method to the baseline classifier using random guess and two broadly evaluated multilabel classification methods in literature: the Green's Function method (Wang et al., 2009) and the Sylvester equation (SMSE) method (Chen et al., 2008).

We also compare the proposed method against two multilabel classification methods designed to study drug resistance in HIV-1: the classifier chain (CC) method and its ensemble version (Read et al., 2011; Heider et al., 2013; denoted as the ensemble of classifier chain [ECC] method). Finally, we also compare our method to two recent multi-instance classification methods: the MTL method (Yuan et al., 2016) designed to study general drug resistance study and the deep multi-instance multilabel (MIML) method (Feng and Zhou, 2017) designed to study general multi-instance data. The Green's Function method and the SMSE methods are implemented following their original articles in Wang et al. (2009) and Chen et al. (2008), respectively, where the parameters are set to the suggested values. The CC method is implemented with logistic regression, where the chaining order for the CC method is 3TC $\rightarrow$ *ABC* $\rightarrow$ *AZT* $\rightarrow$ *d4T* $\rightarrow$ *ddI* as suggested in Heider et al. (2013). Following Heider et al. (2013) and Riemenschneider et al. (2016), we implement the ECC method using both random forests and logistic regression as base classifiers, which are denoted as ''ECC-RF'' and ''ECC-LR,'' respectively. The MTL method and the deep MIML method are implemented using the code published by the respective authors. The overall resistance prediction performances of the compared methods are reported in Table 1.

The comparison results in Table 1 show that the ℓ_1-norm version of the proposed method consistently outperforms all competing methods in terms of all the three performance metrics, sometime very

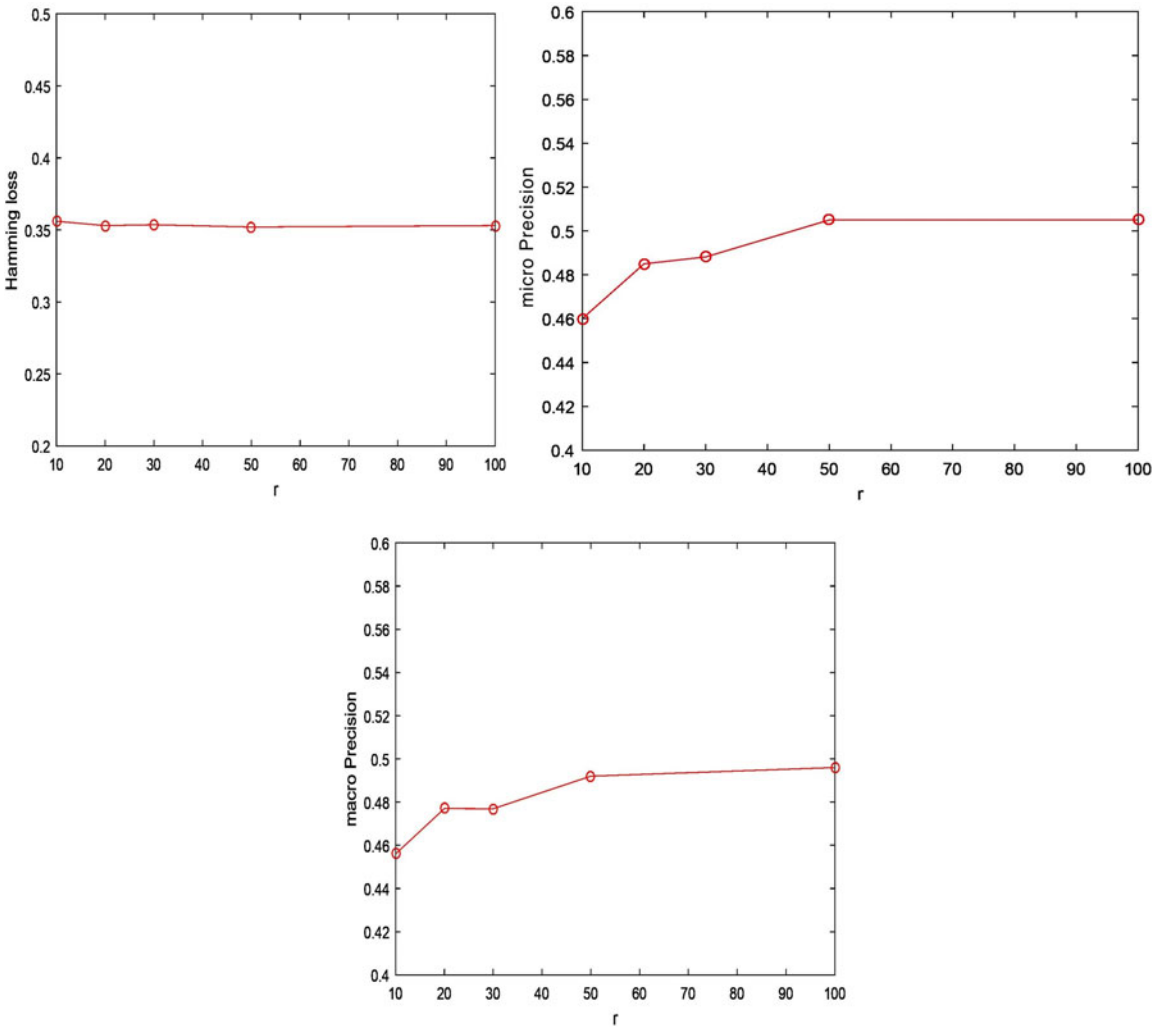

FIG. 3. Multilabel classification performance of the proposed method on the HIV-1 drug resistance data with respect to r (the dimensionality of $\mathbf{W}_k$).

significantly. The squared ℓ_2-norm version of our new method is, as expected, not as effective as its counterpart using the ℓ_1-norm distance, but it still provides adequate performance compared to the other methods in Table 1.

4.4. A case study

We explore the learned distances by our method between RT sequence pairs and compared them with the Euclidean distances for the same RT sequence pairs. The distance between two RT sequences by our method is defined as the sum of the two learned RSSDs: for the k-th class, the pairwise distance between sequence $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ is the sum of $D(C_k, \mathbf{x}_i)$ and $D(C_k, \mathbf{x}_{i'})$. Because we learn a distance metric and SCs for each class, this distance is class dependent. Under this definition, the distances given by our method between sample pairs that belong to the same class are expected to be small and those between sample pairs not belonging to the same class are expected to be large. Using the learned class specific metrics and SCs, we compute the pairwise distances between the RT sequences for every class (nucleoside analog), which are plotted in Figure 4. The Euclidean distances are also plotted for comparison.

To demonstrate the effectiveness of the proposed method, we study the distances between two example RT sequences, which are listed at the top of Figure 4. These two RT sequences are known to be resistant to all five nucleoside analogs. As a result, the pairwise distance between these two RT sequences is expected to be small. However, as can be seen in top left panel of Figure 4, the Euclidean distance between these two

Table 1. Performance of the Compared Methods by Standard Fivefold Cross Validations

Compared methods	*Hamming loss (↓)*	*Micro Precision (↑)*	*Macro Precision (↑)*
Random guess	0.632 ± 0.160	0.276 ± 0.061	0.171 ± 0.051
Green's	0.450 ± 0.040	0.319 ± 0.046	0.241 ± 0.033
SMSE	0.385 ± 0.020	0.402 ± 0.032	0.241 ± 0.020
CC	0.302 ± 0.028	0.467 ± 0.046	0.434 ± 0.037
ECC-LR	0.313 ± 0.014	0.481 ± 0.011	0.442 ± 0.012
ECC-RF	0.301 ± 0.005	0.476 ± 0.020	0.461 ± 0.021
MTL	0.382 ± 0.010	0.475 ± 0.021	0.461 ± 0.010
Deep MIML	0.315 ± 0.010	0.478 ± 0.042	0.474 ± 0.022
Ours-ℓ_2^2	0.322 ± 0.015	0.505 ± 0.040	0.492 ± 0.050
Ours-ℓ_1	**0.282 ± 0.007**	**0.518 ± 0.012**	**0.527 ± 0.013**

Where "↓" means that smaller is better, and "↑" means that bigger is better.

CC, classifier chain; ECC, ensemble of classifier chain; MIML, multi-instance multilabel; MTL, multitask learning; SMSE, Sylvester equation.

RT sequences is ranked at the 1855-th smallest distance among all pairwise Euclidean distances, which is not in accordance with the clinical evidences. In contrast, we can see that the pairwise distances between these RT sequences computed by our learned RSSDs for the five classes are small, which are at the 138-th smallest distance for *3TC*, the 525-th smallest distance for *ABC*, the 574-th smallest distance for *AZT*, the 406-th smallest distance for *d4T*, and 678-th smallest distance for *ddI*, respectively. This observation clearly demonstrates that the learned distances by our new methods can better capture the relationships between data samples in terms of class membership.

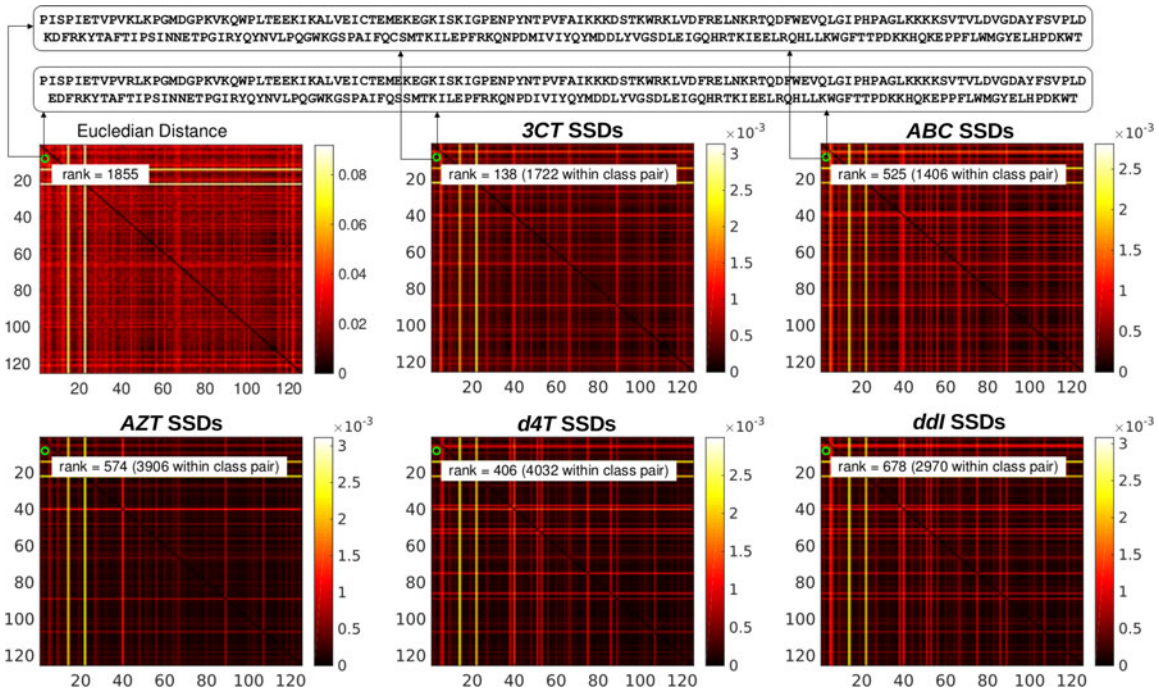

FIG. 4. Exploration of the learned sample-to-sample distance between RT sequence pairs for each class. Top panel: The two RT sequences (with known drug resistance) we are comparing; Top Left Heatmap: the Euclidean distances between RT sequence pairs. Remaining Heatmaps: the learned sample-to-sample distances between RT sequence pairs for each of the five classes. We can see that the sample-to-sample distance between the two RT sequences in the top panel for 3CT nucleoside analog is ranked as the 138-th smallest pairwise distance among all 1722 RT sequence pairs. Compared to the Euclidean distance, which is ranked as 1855-th smallest distance, the pairwise distance computed by the projection and significance coefficients learned for this class is more clinically meaningful. 3TC, Lamivudine; *ABC*, Abacavir; *AZT*, Zidovudine; *d4T*, Stavudine; *ddI*, Didanosine; SSD, sample specific distance.

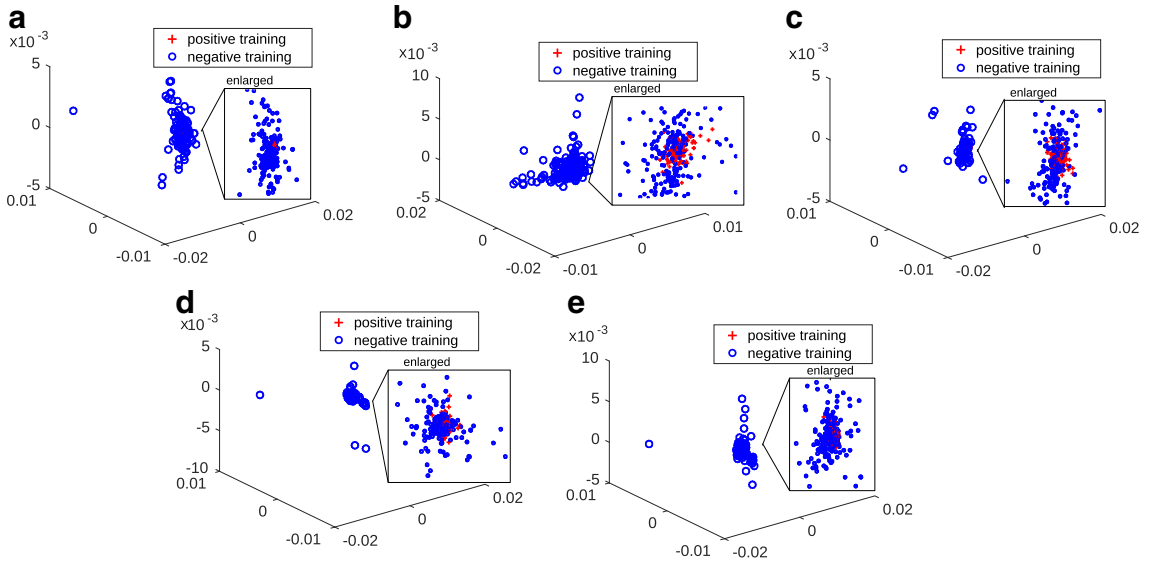

FIG. 5. Projected data samples through the learned projection matrix $\mathbf{W}_k$. The RT sequences belonging to a target class are close in the respective learned subspaces. The regions containing positive data are enlarged for better view. **(a)** *3TC*, **(b)** *ABC*, **(c)** *AZT*, **(d)** d4T, **(e)** *ddI*.

4.5. The discriminative capability of the learned RSSDs

From the subspace learning perspective, the ISD in Equation (6) can be written as

$$D(C_k, \mathbf{x}_{i'}) = \sum_{\mathbf{x}_i \in C_k} \left\| \mathbf{W}_k^T (\mathbf{x}_i - \mathbf{x}_{i'}) s_{ik} \right\|_1, \tag{45}$$

which is the learned C2S distance in the projected lower r-dimensional subspace, where the projection is implemented by $\mathbf{W}_k$ for a given class. Now we study the geometric distributions of the data samples in the subspaces through the learned projections. The results are reported in Figure 5. In this study, the data points

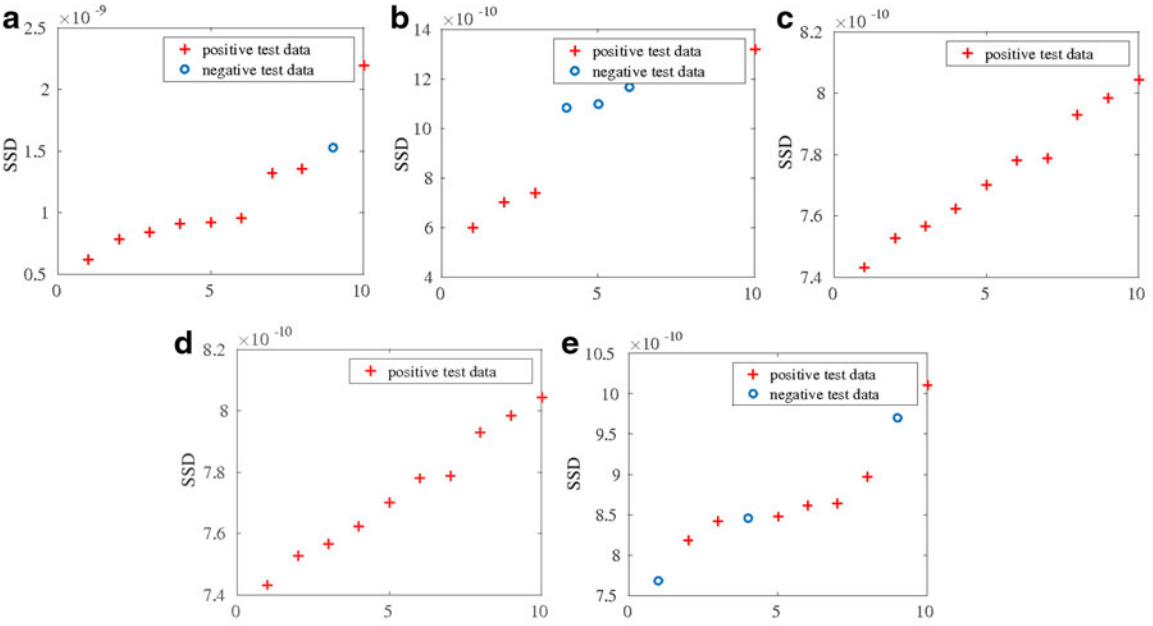

FIG. 6. True labels of top 10 data with smallest RSSDs for each of the five classes. **(a)** *3TC*, **(b)** *ABC*, **(c)** *AZT*, **(d)** *d4T*, **(e)** *ddI*.

in the r-dimensional projected subspaces are mapped into the 3-dimensional space through principal component analysis (Jolliffe, 1986) for the visualization purpose.

From Figure 5 we can see that the proposed method works very well in terms of maximizing data separability. In each of the five projected subspaces, data samples belonging to any one of target classes are close to one another, while those not belonging to the same class are far away from each other. This observation concretely suggests that the learned projected matrices $\mathbf{W}_k$, thereby the learned distance metrics $\mathbf{M}_k$, are able to capture the intrinsic data representations for each class and distinguish between useful and useless features.

To further evaluate the discriminative ability of the proposed method, we plot the top 10 shortest RSSDs among test data for every class in Figure 6. From the results we can see that, shown by red crosses in each figure, most data samples with the smallest RSSDs indeed belong to the target class. This observation once again suggests that the proposed RSSDs with optimized distance metrics and instance-specific SCs are good criteria for drug resistance prediction.

5. CONCLUSIONS

In this study, we proposed a novel RSSD method for multilabel classification. To learn the parameters of the proposed RSSDs, we formulated a learning objective that maximizes the ratio of the summations of a number of ℓ_1-norm distances, which is difficult to solve in general. To solve this problem we derived a new efficient iterative algorithm with rigorously proved convergence. The promising experimental results have demonstrated the effectiveness of our new method for identifying HIV-1 drug resistances.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

This work was partially supported by the National Science Foundation under the grants of IIS-1652943 and IIS-1849359.

REFERENCES

Chen, G., Song, Y., Wang, F., et al. 2008. Semi-supervised multi-label learning by solving a Sylvester equation. Proceedings of the 2008 SIAM International Conference on Data Mining, 410–419.

Ding, C., Zhou, D., He, X., et al. 2006. R1-PCA: Rotational invariant l1-norm principal component analysis for robust subspace factorization. Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 281–288.

Fan, K. 1950. On a theorem of weyl concerning eigenvalues of linear transformations II. *Proc. Natl. Acad. Sci. USA*. 36, 31–35.

Feng, J., and Zhou, Z.-H. 2017. Deep MIML network. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA.

Fukunaga, K. 2013. *Introduction to Statistical Pattern Recognition*. Amsterdam: Elsevier.

Gönen, M., and Margolin, A.A. 2014. Drug susceptibility prediction against a panel of drugs using kernelized bayesian multitask learning. *Bioinformatics*. 30, i556–i563.

Han, F., Wang, H., and Zhang, H. 2018. Learning integrated holism-landmark representations for long-term loop closure detection. The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18).

Heider, D., Senge, R., Cheng, W., et al. 2013. Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction. *Bioinformatics*. 29, 1946–1952.

Heider, D., Verheyen, J., and Hoffmann, D. 2010. Predicting bevirimat resistance of HIV-1 from genotype. *BMC Bioinformatics*. 11, 37.

Hepler, N.L., Scheffler, K., Weaver, S., et al. 2014. IDEPI: Rapid prediction of HIV-1 antibody epitopes and other phenotypic features from sequence data using a flexible machine learning platform. *PLoS Comput. Biol*. 10, e1003842.

Jenatton, R., Obozinski, G., and Bach, F. 2010. Structured sparse principal component analysis. Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy.

Jolliffe, I.T. 1986. Principal component analysis and factor analysis, 115–128. *In* Jolliffe, I.T., ed. *Principal Component Analysis*. Springer, New York, NY.

Ke, Q., and Kanade, T. 2005. Robust l/sub 1/norm factorization in the presence of outliers and missing data by alternative convex programming. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, 1, 739–746.

Kwak, N. 2008. Principal component analysis based on l1-norm maximization. *IEEE Trans. Pattern. Anal*. 30, 1672–1680.

Kyte, J., and Doolittle, R.F. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol*. 157, 105–132.

Lewis, D.D., Yang, Y., Rose, T.G., et al. 2004. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res*. 5, 361–397.

Liu, K., Brand, L., Wang, H., et al. 2019a. Learning robust distance metric with side information via ratio minimization of orthogonally constrained ℓ_{21}-norm distances. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19).

Liu, K., Wang, H., Han, F., et al. 2019b. Visual place recognition via robust l2-norm distance based holism and landmark integration. Proceedings of the AAAI Conference on Artificial Intelligence.

Liu, K., Wang, H., Nie, F., et al. 2018. Learning multi-instance enriched image representations via non-greedy ratio maximization of the ℓ_1-norm distances. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 7727–7735.

Liu, Y., Gao, Q., Miao, S., et al. 2017. A non-greedy algorithm for l1-norm LDA. *IEEE Trans. Image Process.* 26, 684–695.

Pennings, P.S. (2012). Standing genetic variation and the evolution of drug resistance in HIV. *PLoS Comput. Biol.* 8, e1002527.

Read, J., Pfahringer, B., Holmes, G., et al. 2011. Classifier chains for multi-label classification. *Mach. Learn*. 85, 333–359.

Rhee, S.-Y., Gonzales, M. J., Kantor, R., et al. 2003. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res*. 31, 298–303.

Rhee, S.-Y., Taylor, J., Wadhera, G., et al. 2006. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proc. Natl. Acad. Sci. USA*. 103, 17355–17360.

Riemenschneider, M., Senge, R., Neumann, U., et al. 2016. Exploiting HIV-1 protease and reverse transcriptase cross-resistance information for improved drug resistance prediction by means of multi-label classification. *Biodata. Min.* 9, 10.

Smyth, R.P., Davenport, M. P., and Mak, J. 2012. The origin of genetic diversity in HIV-1. *Virus. Res.* 169, 415–429.

Sun, W., and Yuan, Y.-X. 2006. *Optimization Theory and Methods: Nonlinear Programming*, Volume 1. Springer Science & Business Media, New York.

Wang, H., Deng, C., Zhang, H., et al. 2016. Drosophila gene expression pattern annotations via multi-instance biological relevance learning. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16).

Wang, H., Ding, C., and Huang, H. 2010a. Multi-label linear discriminant analysis. 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 126–139.

Wang, H., Ding, C.H., and Huang, H. 2010b. Multi-label classification: Inconsistency and class balanced k-nearest neighbor. Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10).

Wang, H., Huang, H., and Ding, C. 2009. Image annotation using multi-label correlated green's function. 2009 IEEE 12th International Conference on Computer Vision, 2029–2034.

Wang, H., Huang, H., and Ding, C. 2010c. Multi-label feature transform for image classifications. 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 793–806.

Wang, H., Huang, H., and Ding, C. 2013a. Function–function correlated multi-label protein function prediction over interaction networks. *J. Comput. Biol*. 20, 322–343.

Wang, H., Huang, H., and Ding, C. 2015. Correlated protein function prediction via maximization of data-knowledge consistency. *J. Comput. Biol.* 22, 546–562.

Wang, H., Huang, H., Kamangar, F., et al. (2011a). Maximum margin multi-instance learning. NIPS, 1–9.

Wang, H., Nie, F., and Huang, H. 2011b. Learning instance specific distance for multi-instance classification. Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence.

Wang, H., Nie, F., and Huang, H. 2012. Robust and discriminative distance for multi-instance learning. 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI.

Wang, H., Nie, F., and Huang, H. 2013b. Robust and discriminative self-taught learning. *ICML.* 28, 298–306.

Wang, H., Nie, F., and Huang, H. 2014. Robust distance metric learning via simultaneous ℓ_1-norm minimization and maximization. *ICML.* 32, 1836–1844.

Wang, H., Nie, F., Huang, H., et al. 2011c. Learning frame relevance for video classification. Proceedings of the 19th ACM International Conference on Multimedia, Scottsdale, AZ, 1345–1348.

Wright, J., Ganesh, A., Rao, S., et al. 2009. Robust principal component analysis: Exact recovery of corrupted. *NIPS* 116.

Wright, S.J., and Nocedal, J. 1999. Numerical optimization. *Science*. 35, 7.

Yang, H., Liu, K., Wang, H., et al. 2019. Learning strictly orthogonal *p*-order nonnegative laplacian embedding via smoothed iterative reweighted method. *IJCAI*. 2019, 4040–4046.

Yuan, H., Paskov, I., Paskov, H., et al. 2016. Multitask learning improves prediction of cancer drug sensitivity. *Sci. Rep.* 6, 31619.

Address correspondence to:
Dr. Hua Wang
Department of Computer Science
Colorado School of Mines
Brown Building 280F
1610 Illinois Street
Golden, CO 80401

E-mail: huawangcs@gmail.com