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

Lodewijk Brand, Xue Yang, Kai Liu, Saad Elbeleidy, Hua Wang[(✉)], and Hao Zhang

Department of Computer Science, Colorado School of Mines, Golden, CO 80401, USA
{lbrand,selbeleidy,hzhang}@mines.edu, edyxueyx@gmail.com,
liukaizhijia@gmail.com, huawangcs@gmail.com

Abstract. Acquired immunodeficiency syndrome (AIDS) is a syndrome caused by the human immunodeficiency virus (HIV). During the progression of AIDS, a patient's the 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 multi-label classification problem. Given this multi-class relationship, traditional single-label classification methods usually fail to effectively identify the drug resistances that may develop after a particular virus mutation. In this paper, we propose a novel multi-label Robust Sample Specific Distance (RSSD) method to identify multi-class HIV drug resistance. Our method is novel in that it can illustrate the relative strength of the drug resistance of a reverse transcriptase sequence against a given drug nucleoside analogue 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, non-greedy, iterative algorithm with rigorously proved convergence. Our new method has been verified on a public HIV-1 drug resistance data set with over 600 RT sequences and five nucleoside analogues. We compared our method against other state-of-the-art multi-label classification methods and the experimental results have demonstrated the effectiveness of our proposed method.

Keywords: Human immunodeficiency virus · Drug resistance · Multi-label classification

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 51–67, 2019.
https://doi.org/10.1007/978-3-030-17083-7_4

1 Introduction

According to estimations by the World Health Organization, around 35 million people suffer from the Human immunodeficiency virus (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 stem the progression of HIV and extend the lifespan 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 [24]. The mutation of HIV-1 during antiretroviral treatments can lead to a phenomenon called "cross-resistance" [7,23]. 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 the 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. In order to address this problem, it is important that we develop automatic methods that can associate genetic strains of HIV to their corresponding drug resistances.

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 [21]. Furthermore, 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 [19]. 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 by using short sequence information of the viral genotype, such as the genetic sequence of the viral reverse transcriptase (RT). For example, Rhee *et al.* [22] 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 [9]. Additionally, a Bayesian algorithm that combines kernel-based nonlinear dimensionality reduction and binary classification has been proposed to predict drug susceptibility of HIV within a multi-task learning framework [5]. 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 [7,23]. To tackle this difficulty, in this paper we propose to solve the problem of HIV-1 drug resistance prediction as a *multi-label classification* problem.

Multi-label classification is an emerging research topic in machine learning driven by the advances of modern technologies in recent years [27–32,39]. As a generalization of traditional single-label classification that requires every data sample to belong to one and only one class, multi-label classification relaxes this restriction and allows a 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 multi-label classification problems are interdependent on one another. Although the labeling relaxation in multi-label classification problems have brought a number of successes in a variety of real-world applications [29,30,32], it also causes labeling ambiguity that inevitably complicates the problem [27,28]. 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 multiples classes to train a classifier for a given class [27,28]. 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 [27,29], which is equivalent to assume that every data sample contributes equally to a trained classification model [28]. However, this is not true in most real-world multi-label classification problems. For example, some RT sequences natively resist against a certain drug. On the other hand, the same RT sequences can develop resistances against other drugs through mutations, which is assumed to be not as strong as native resistances. Simply put, in order to create an effective multi-label 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 [28].

In this paper we propose a novel Robust Sample Specific Distance (RSSD) for multi-label data to predict HIV-1 drug resistance, which, as illustrated in Fig. 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 classes. 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 previous works [6,15], we derive an efficient iterative algorithm with theoretically guaranteed convergence, which, different from our previous works [35,37], is a *non-greedy* algorithm such that it has a better chance to find the optima of the proposed objective. In addition, as an important theoretical contribution of this paper, our new algorithm solves the general optimization problem that maximizes the ratio of the summations of the ℓ_1-norm distances in a non-greedy way, which can find many applications to improve a number of machine learning models. We applied our new method to predict the HIV-1 drug resistance on a public benchmark data set. The experimental results have shown that our new RSSD method outperforms other state-of-the-art competing methods.

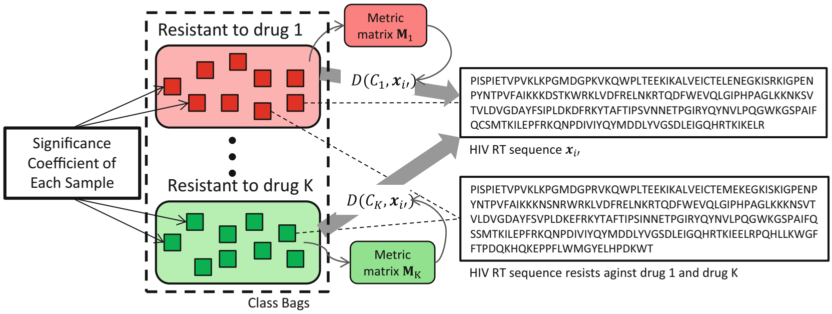
Resistant to drug 1
Metric matrix $\mathbf{M}_1$
$D(C_1, \boldsymbol{x}_{i'})$
PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTELENEGKISRKIGPEN
PYNTPVFAIKKKDSTKWRKLVDFRELNKRTQDFWEVQLGIPHPAGLKKNKSV
TVLDVGDAYFSIPLDKDFRKYTAFTIPSVNNETPGIRYQYNVLPQGWKGSPAIF
QCSMTKILEPFRKQNPDIVIYQYMDDLYVGSDLEIGQHRTKIKELR
HIV RT sequence $\boldsymbol{x}_{i'}$
Significance Coefficient of Each Sample
Resistant to drug K
$D(C_K, \boldsymbol{x}_{i'})$
PISPIETVPVKLKPGMDGPRVKQWPLTEEKIKALVEICTEMEKEGKISKIGPENP
YNTPVFAIKKKNSNRWRKLVDFRELNKRTQDFWEVQLGIPHPAGLKKKNSVT
VLDVGDAYFSVPLDKEFRKYTAFTIPSINNETPGIRYQYNVLPQGWKGSPAIFQ
SSMTKILEPFRKQNPDIVIYQYMDDLYVGSDLEIGQHRTKIEELRPQHLLKWGF
FTPDQKHQKEPPFLWMGYELHPDKWT
Metric matrix $\mathbf{M}_K$
HIV RT sequence resists aginst drug 1 and drug K
Class Bags

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 analogue). 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, *i.e.*, it is a multi-label data sample. Ideally, the learned Significance Coefficients 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. (Color figure online)

2 Learning Robust Sample Specific Distances (RSSDs) for Multi-label Classification

In this section, we first formalize the problem of predicting the drug resistance of HIV-1. Then we derive a novel RSSD to solve the problem following previous works [26,33–35,38] that solve multi-instance problems.

Throughout this paper, 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 $\|\mathbf{M}\|_{\text{F}}$ and we define its ℓ_1-norm as $\|\mathbf{M}\|_1 = \sum_i \sum_j |m_{ij}|$. The trace of $\mathbf{M}$ is defined as $\mathbf{tr}\ (\mathbf{M}) = \sum_i m_{ii}$.

In a multi-label 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 analogues). Here $\mathbf{x}_i \in \Re^d$, 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 analogues (drug variants) a HIV-1 RT sequence is resistant to.

2.1 The Class-to-Sample (C2S) Distance

To learn the distance from a class to a data sample, we first represent each class as a bag that consists of all samples that belong to this class, *i.e.*, $C_k = \{\mathbf{x}_i | i \in \pi_k\}$, where $\pi_k = \{i | Y_{ik} = 1\}$ is the set of indices of all training samples that belong to the k-th class. The number of samples in C_k is denoted as m_k, *i.e.*, $|C_k| = m_k$. Note that, in a single-label classification problem, a data sample precisely belongs to one and only one class at a time. It follows that $\sum_{i=1}^{K} Y_{ik} = 1$ and $C_k \cap C_l = \emptyset$ $(\forall k \neq l)$. In contrast, in a multi-label classification problem, a data sample may belong to more than one class at the same time. It can happen that $C_k \cap C_l \neq \emptyset$ $(\exists k \neq l)$, *i.e.*, different class bags may overlap and an individual data sample $\mathbf{x}_i$ may appear in multiple class bags.

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

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

We then compute the 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 $\mathbf{x}_{i'}$ as following:

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

2.2 Parameterized C2S Distance

Because the C2S distance in Eq. (2) does not take into account the resistance strength against a certain nucleoside analogue, 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 in a multi-label classification problem [27,28], some samples in a class may characterize that particular class more strongly than the others from the statistical point of view. For example [23], 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. We must capture both of these in order for our method to be effective. As a result, we should assign less weight to less resistant RT sequences when determining whether to apply the "resistant" label to a query viral RT sequence.

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

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

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 viral RT sequences against the target nucleoside analogues during the learning process.

2.3 Parameterized C2S Distance Refined by Class Specific Distance Metrics

The RSSD defined in Eq. (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 in order 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:

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

Because the class-specific distance metric $\mathbf{M}_k$ is a positive definite matrix, 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 $\mathbf{W}_k^T \mathbf{W}_k = \mathbf{I}$. Here we can also reasonably assume that $d > r$, because $\mathbf{M}_k = \mathbf{W}_k \mathbf{W}_k^T$ is a $d \times d$ matrix and its maximum rank is d. Thus we can rewrite Eq. (4) as follows:

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

A critical problem of $D(C_k, \mathbf{x}_{i'})$ defined in Eq. (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 [2,37]. Due to the cross-resistance phenomenon [7], this problem is particularly critical for identifying HIV-1 drug resistance. To promote the robustness of $D(C_k, \mathbf{x}_{i'})$ against outliers, following many previous works [11,12,17,18,36,37,40], we define it using the ℓ_1-norm distance as follows:

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

which we call the proposed *Robust Sample Specific Distance (RSSD)*.

To use RSSD defined in Eq. (6), we need to learn two sets of parameters s_{ik} and $\mathbf{W}_k$ for every class. Following the most broadly used machine learning strategy to maximize data discriminativity for classification, such as Fisher's linear discriminant [4], for a given class C_k we simultaneously maximize the overall RSSDs from every class bag to all its non-belonging samples and minimize the overall RSSDs from every class bag to all the samples belonging to that class:

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

Algorithm 1. Algorithm to solve Eq. (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: $v$.
```

Learning the RSSDs by solving Eq. (7) and classifying query viral RT sequences using the adaptive decision boundary method [29], our proposed RSSD method can be used for identifying HIV-1 drug resistance, as well as general multi-label classification problems.

3 An Efficient Solution Algorithm

Our new objective in Eq. (7) maximizes the ratio of the summations of a number of ℓ_1-norm distances, which is obviously not smooth and therefore difficult to solve in general. To solve this challenging optimization problem, we use the optimization method proposed in our previous works in [6,15].

We first turn to solve the following generalized the objective:

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

where Ω is the feasible domain. Next, we propose a simple, yet efficient, iterative framework in Algorithm 1 to solve the objective in Eq. (8). The convergence of Algorithm 1 is rigorously guaranteed by Theorem 1. Due to space limit, the proofs of all the theorems in this paper are provided in the extended journal version of this paper.

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* $\forall t > \hat{t}$ $\frac{h(v^t)}{m(v^t)} - \frac{h(v^{t-1})}{m(v^{t-1})} < \delta$.

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 Eq. (7) as:

$$F(\mathbf{W}_k) = H(\mathbf{W}_k) - \lambda^t M(\mathbf{W}_k) \geq 0, \tag{9}$$

where λ^t is computed by

$$\lambda^t = \frac{\sum_{\mathbf{x}_i' \notin C_k} \sum_{\mathbf{x}_i \in C_k} \left\| (\mathbf{W}_k^{t-1})^T (\mathbf{x}_i - \mathbf{x}_{i'}) s_{ik} \right\|_1}{\sum_{\mathbf{x}_i' \in C_k} \sum_{\mathbf{x}_i \in C_k} \left\| (\mathbf{W}_k^{t-1})^T (\mathbf{x}_i - \mathbf{x}_{i'}) s_{ik} \right\|_1}. \tag{10}$$

In Eq. (10), $\mathbf{W}_k^{t-1}$ denotes the projection matrix in the $(t-1)$-th iteration. Here, we define the following:

$$\begin{aligned} H(\mathbf{W}_k) &= \sum_{\mathbf{x}_i' \notin C_k} \sum_{\mathbf{x}_i \in C_k} \left\| \mathbf{W}_k^T \left(\mathbf{x}_i - \mathbf{x}_{i'} \right) s_{ik} \right\|_1 , \\ M(\mathbf{W}_k) &= \sum_{\mathbf{x}_i' \in C_k} \sum_{\mathbf{x}_i \in C_k} \left\| \mathbf{W}_k^T \left(\mathbf{x}_i - \mathbf{x}_{i'} \right) s_{ik} \right\|_1 . \end{aligned} \tag{11}$$

Now we need solve the problem in Eq. (9), for which we first introduce the following two lemmas:

Lemma 1. *[16, 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. *[10, 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:

$$L(\mathbf{W}_k, \mathbf{W}_k^{t-1}) = K(\mathbf{W}_k) - \lambda^t N(\mathbf{W}_k), \tag{12}$$

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

$$\begin{aligned} K(\mathbf{W}_k) &= \sum_{g=1}^{r} \mathbf{w}_g^T \mathbf{B} \, \text{sign} \left(\mathbf{B}^T \mathbf{w}_g^{t-1} \right) , \\ N(\mathbf{W}_k) &= \frac{1}{2} \sum_{g=1}^{r} \mathbf{w}_g^T \mathbf{A}_g \mathbf{w}_g + \left(\mathbf{w}_g^{t-1} \right)^T \mathbf{A}_g \mathbf{w}_g^{t-1} . \end{aligned} \tag{13}$$

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:

$$\begin{aligned} \mathbf{B} &= \left[\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}, \bar{\mathbf{x}}_2 - \bar{\mathbf{x}}, \cdots, \bar{\mathbf{x}}_n - \bar{\mathbf{x}} \right] , \\ \mathbf{A}_g &= \sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \frac{\left(\mathbf{x}_j - \bar{\mathbf{x}}_i \right) \left(\mathbf{x}_j - \bar{\mathbf{x}}_i \right)^T}{\left| \left(\mathbf{w}_g^{t-1} \right)^T \left(\mathbf{x}_j - \bar{\mathbf{x}}_i \right) \right|} , \end{aligned} \tag{14}$$

and sign(x) is the sign function.

Then, using the definition of $L(\mathbf{W}_k, \mathbf{W}_k^{t-1})$ in Eq. (12) and Lemmas 1 and 2, we can prove the following theorem:

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

$$L(\mathbf{W}_k, \mathbf{W}_k^{t-1}) \leq F(\mathbf{W}_k). \tag{15}$$

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

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

```
Input: $\mathbf{W}_k^{t-1}$ and Armijo parameter $0 < \beta < 1$.
1. Calculate $\lambda^k$ by Eq. (10).
2. Calculate the subgradient
$\mathbf{G}^{k-1} = \partial L(\mathbf{W}_k^{t-1}, \mathbf{W}_k^{t-1}) = \mathbf{B}\,\mathrm{sign}\left(\mathbf{B}^T\mathbf{W}_k^{t-1}\right) - \lambda^k\left[\mathbf{A}_1\mathbf{w}_1, \mathbf{A}_2\mathbf{w}_2, \cdots, \mathbf{A}_r\mathbf{w}_r\right]$.
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 Eq. (9).
    6. $t = t + 1$.
Output: $\mathbf{W}_k^k$.
```

Algorithm 3. Algorithm for non-greedy 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 $\lambda^t$ by Eq. (10).
    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: $\mathbf{W}$.
```

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 Eq. (7), $\mathbf{W}_k^\star$ satisfies $L(\mathbf{W}_k^\star, \mathbf{W}_k^{t-1}) \geq L(\mathbf{W}_k^{t-1}, \mathbf{W}_k^{t-1}) = 0$. Then, we have:

$$F(\mathbf{W}_k^\star) \geq L(\mathbf{W}_k^\star, \mathbf{W}_k^{t-1}) \geq L(\mathbf{W}_k^{t-1}, \mathbf{W}_k^{t-1}) = F(\mathbf{W}_k^{t-1}) = 0. \qquad (16)$$

Lemma 1 and Eq. (16) indicate that the solution of the objective function in Eq. (9) can be transformed to solve the objective function $L(\mathbf{W}_k, \mathbf{W}_k^{t-1}) \geq 0$, which can be easily solved by the projected subgradient method with Armijo line search [25]. 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, *i.e.* $(\mathbf{W}_k^t)^T(\mathbf{W}_k^t) = \mathbf{I}$. Algorithm 2 summarizes the algorithm to solve the objective in Eq. (9).

Finally, based on Algorithm 2, we can derive a simple yet efficient iterative algorithm as summarized in Algorithm 3 to solve our objective in Eq. (7) when s_{ik} is fixed. In addition, Theorem 3 indicates that our proposed Algorithm 3 monotonically increase 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.

Theorem 3. *If $\mathbf{W}_k^t$ is the solution of the objective function in Eq. (9) and satisfies $(\mathbf{W}_k^t)^T(\mathbf{W}_k^t) = \mathbf{I}$, then we have $\mathcal{J}(\mathbf{W}_k^t) \geq \mathcal{J}(\mathbf{W}_k^{t-1})$.*

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

3.2 Fixing $\mathbf{W}_k$ to Solve s_{ik}

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

$$\max \frac{\sum_{\mathbf{x}_i' \notin C_k} \sum_{\mathbf{x}_i \in C_k} s_{ik} d_{ii'k}}{\sum_{\mathbf{x}_i' \in C_k} \sum_{\mathbf{x}_i \in C_k} s_{ik} d_{ii'k}}, \quad s.t.\ s_{ik} \geq 0. \tag{17}$$

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:

$$\max \frac{\sum_{\mathbf{x}_i \notin C_k} s_{ik} d_{ik}^w}{\sum_{\mathbf{x}_i \in C_k} s_{ik} d_{ik}^b}, \quad s.t.\ s_{ik} \geq 0. \tag{18}$$

Again, to solve Eq. (18), to Step 3 in Algorithm 1, we solve the following optimization problem:

$$\max \sum_{\mathbf{x}_i \in C_k} s_{ik} d_{ik}^w - \lambda \sum_{\mathbf{x}_i \in C_k} s_{ik} d_{ik}^b, \quad s.t.\ s_{ik} \geq 0, \tag{19}$$

where λ is computed as Eq. (10) in the t-th iteration.

Define that $d_{ik} = d_{ik}^w - \lambda d_{ik}^b$, we can rewrite the optimization problem in Eq. (19) as:

$$\max \sum_{\mathbf{x}_i \in C_k} s_{ik} d_{ik}, \quad s.t.\ s_{ik} \geq 0, \tag{20}$$

The problem in Eq. (20) can be decoupled to solve the following subproblems separately for each $\mathbf{x}_i \in C_k$:

$$\max\ s_{ik} d_{ik}, \quad s.t.\ s_{ik} \geq 0, \tag{21}$$

which is a convex linear programming problem [41] and can be solved efficiently by many off-the-shelf solution algorithms [41]. By inserting the solution to Eq. (21) after Step 3 of Algorithm 3, we can finally solve our objective in Eq. (7), which is equivalent to perform 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 [22], 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 analogues: Lamivudine (3TC), Abacavir (*ABC*), Zidovudine (*AZT*), Stavudine (*d4T*) and Didanasine (*ddI*). Following [8], although the Tenofovir (*TDF*) nucleoside analogue is included in this database, it is not used in our study, because the number of the RT sequences resistant to this nucleoside analogue is very low. 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 [7]. We label the viral RT sequences as "resistant" using the same drug-specific IC_{50} ratio cutoff thresholds as in [7], 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 [13] to represent the RT sequences, which has demonstrated good prediction performance in many protein classification studies [8]. 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 [7] we set a fixed window size of 11 and interpolate all hydrophobicity vectors to length 230 using the spline interpolation method [13].

4.1 Comparative Studies

Predicting drug resistance for HIV-1 RT sequences is a multi-label classification problem. Therefore, we evaluate the proposed method by two broadly used multi-label performance metrics [14]: 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 multi-label classification, the 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.

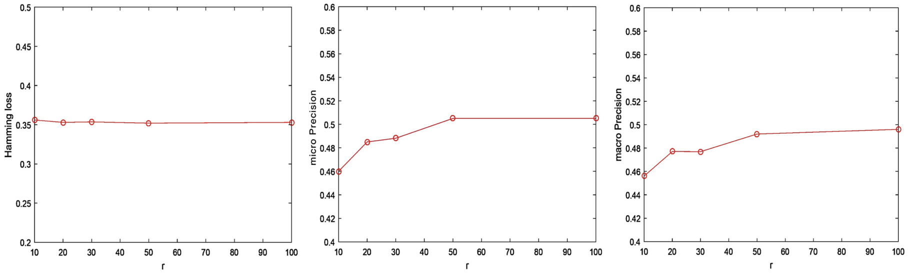

Fig. 2. Multi-label classification performance of the proposed method on the HIV-1 drug resistance data with respect to r (the dimensionality of $\mathbf{W}_k$).

Parameter Selection. The proposed RSSD has only one parameter: the dimensionality r of $\mathbf{W}_k$. Ideally, each class can have its own fine tuned parameter. Although, to reduce the experimental effort, we fix the parameter r across all classes in our studies. We evaluate the impacts of the parameter in a standard 5-fold cross-validation experiment, where we select r in the range from 10 to 100. The classification performance measured by the three aforementioned performance metrics are reported in Fig. 2, when we vary r. 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; this property 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.

Comparative Studies. We use a standard 5-fold 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 Eq. (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 Eq. (5) (denoted as "Ours-ℓ_2^2"). We compare our new method to two broadly evaluated multi-label classification methods in literature: the Green's Function method [29] and the Sylvester Equation (SMSE) method [1]. We also compare the proposed method against two multi-label classification methods designed to study drug resistance in HIV-1: the Classifier Chain (CC) method and its ensemble version [7,20] (denoted as the ECC method). Finally, we also compare our method to two recent multi-instance classification methods: the multi-task learning (MTL) method [42] designed to study general drug resistance study and the deep MIML method [3] designed to study general multi-instance data. The Green's Function method and the Sylvester Equation methods are implemented following their original papers in [29] and [1] 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 → $ABC \rightarrow AZT \rightarrow d4T \rightarrow ddI$ as suggested in [7]. Following [7,23], we implement the ECC method by 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 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 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 when compared to the other methods in Table 1.

4.2 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

Table 1. Performance of the compared methods by standard 5-fold cross validations, where "↓" means that smaller is better and "↑" means that bigger is better.

Compared methods	Hamming loss (↓)	Micro precision (↑)	Macro precision (↑)
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**

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 significance coefficients 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 significance coefficients, we compute the pairwise distances between the RT sequences for every class (nucleoside analogue), which are plotted in Fig. 3. 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 Fig. 3. These two RT sequences are known to be resistant to all five nucleoside analogues. As a result, the pairwise distance between these two RT sequences are expected to be small. However, as can be seen in top left panel of Fig. 3, the Euclidean distance between these two 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 *3CT*, 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.

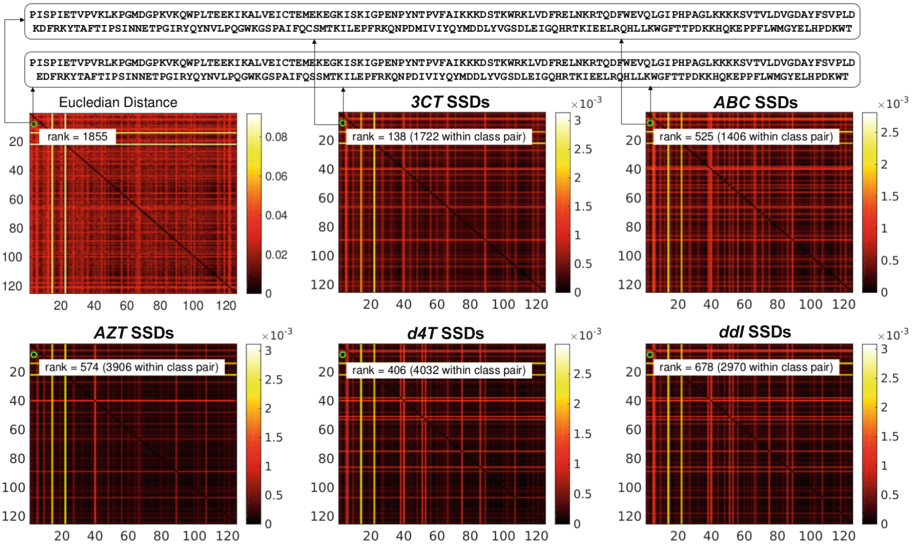

Fig. 3. Exploration of the learned sample-to-sample distance between RT sequence pairs for each class. **Top panel:** The two RT sequences (with known drug resistance) we are comparing; **Top Left Heatmap:** the Euclidean distances between RT sequence pairs. **Remaining Heatmaps:** the learned sample-to-sample 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 analogue 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.

5 Conclusions

In this paper, we proposed a novel RSSD method for multi-label 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; this problem 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.

Acknowledgments. This work was partially supported by National Science Foundation under Grant NSF-IIS 1652943. This research was also partially supported by Army Research Office (ARO) under Grant W911NF-17-1-0447, U.S. Air Force Academy (USAFA) under Grant FA7000-18-2-0016, and the Distributed and Collaborative Intelligent Systems and Technology (DCIST) CRA under Grant W911NF-17-2-0181.

References

1. Chen, G., Song, Y., Wang, F., Zhang, C.: Semi-supervised multi-label learning by solving a sylvester equation. In: SDM, pp. 410–419. SIAM (2008)
2. Ding, C., Zhou, D., He, X., Zha, H.: R1-PCA: rotational invariant L1-norm principal component analysis for robust subspace factorization. In: ICML, pp. 281–288 (2006)
3. Feng, J., Zhou, Z.H.: Deep MIML network. In: AAAI (2017)
4. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Elsevier, Amsterdam (2013)
5. Gönen, M., Margolin, A.A.: Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning. Bioinformatics **30**(17), i556–i563 (2014)
6. Han, F., Wang, H., Zhang, H.: Learning of integrated holism-landmark representations for long-term loop closure detection. In: AAAI Conference on Artificial Intelligence (2018)
7. Heider, D., Senge, R., Cheng, W., Hüllermeier, E.: Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction. Bioinformatics **29**(16), 1946–1952 (2013)
8. Heider, D., Verheyen, J., Hoffmann, D.: Predicting bevirimat resistance of HIV-1 from genotype. BMC Bioinform. **11**(1), 37 (2010)
9. Hepler, N.L., et al.: 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**(9), e1003842 (2014)
10. Jenatton, R., Obozinski, G., Bach, F.: Structured sparse principal component analysis. In: International Conference on Artificial Intelligence and Statistics (2010)
11. Ke, Q., Kanade, T.: Robust L/sub 1/norm factorization in the presence of outliers and missing data by alternative convex programming. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 739–746. IEEE (2005)
12. Kwak, N.: Principal component analysis based on L1-norm maximization. IEEE Trans. Pattern Anal. Mach. Intell. **30**, 1672–1680 (2008)
13. Kyte, J., Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157**(1), 105–132 (1982)
14. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: Rcv1: a new benchmark collection for text categorization research. J. Mach. Learn. Res. **5**, 361–397 (2004)
15. Liu, K., Wang, H., Nie, F., Zhang, H.: Learning multi-instance enriched image representations via non-greedy ratio maximization of the L1-norm distances. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7727–7735 (2018)
16. Liu, Y., Gao, Q., Miao, S., Gao, X., Nie, F., Li, Y.: A non-greedy algorithm for L1-norm LDA. IEEE Trans. Image Process. **26**(2), 684–695 (2017)
17. Nie, F., et al.: New L1-norm relaxations and optimizations for graph clustering. In: AAAI, pp. 1962–1968 (2016)
18. Nie, F., Wang, H., Huang, H., Ding, C.: Unsupervised and semi-supervised learning via ℓ_1-norm graph. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2268–2273. IEEE (2011)
19. Pennings, P.S.: Standing genetic variation and the evolution of drug resistance in HIV. PLoS Comput. Biol. **8**(6), e1002527 (2012)

20. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Mach. Learn. **85**(3), 333–359 (2011)
21. Rhee, S.Y., Gonzales, M.J., Kantor, R., Betts, B.J., Ravela, J., Shafer, R.W.: Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res. **31**(1), 298–303 (2003)
22. Rhee, S.Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D.L., Shafer, R.W.: Genotypic predictors of human immunodeficiency virus type 1 drug resistance. Proc. Natl. Acad. Sci. **103**(46), 17355–17360 (2006)
23. Riemenschneider, M., Senge, R., Neumann, U., Hüllermeier, E., Heider, D.: Exploiting HIV-1 protease and reverse transcriptase cross-resistance information for improved drug resistance prediction by means of multi-label classification. BioData Min. **9**(1), 10 (2016)
24. Smyth, R.P., Davenport, M.P., Mak, J.: The origin of genetic diversity in HIV-1. Virus Res. **169**(2), 415–429 (2012)
25. Sun, W., Yuan, Y.X.: Optimization Theory and Methods: Nonlinear Programming, vol. 1. Springer, Heidelberg (2006). https://doi.org/10.1007/b106451
26. Wang, H., Deng, C., Zhang, H., Gao, X., Huang, H.: Drosophila gene expression pattern annotations via multi-instance biological relevance learning. In: AAAI, pp. 1324–1330 (2016)
27. Wang, H., Ding, C., Huang, H.: Multi-label linear discriminant analysis. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6316, pp. 126–139. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15567-3_10
28. Wang, H., Ding, C.H., Huang, H.: Multi-label classification: inconsistency and class balanced k-nearest neighbor. In: AAAI (2010)
29. Wang, H., Huang, H., Ding, C.: Image annotation using multi-label correlated green's function. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 2029–2034. IEEE (2009)
30. Wang, H., Huang, H., Ding, C.: Multi-label feature transform for image classifications. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 793–806. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_57
31. Wang, H., Huang, H., Ding, C.: Function-function correlated multi-label protein function prediction over interaction networks. J. Comput. Biol. **20**(4), 322–343 (2013)
32. Wang, H., Huang, H., Ding, C.: Correlated protein function prediction via maximization of data-knowledge consistency. J. Comput. Biol. **22**(6), 546–562 (2015)
33. Wang, H., Huang, H., Kamangar, F., Nie, F., Ding, C.H.: Maximum margin multi-instance learning. In: Advances in Neural Information Processing Systems, pp. 1–9 (2011)
34. Wang, H., Nie, F., Huang, H.: Learning instance specific distance for multi-instance classification. In: AAAI, vol. 2, p. 6 (2011)
35. Wang, H., Nie, F., Huang, H.: Robust and discriminative distance for multi-instance learning. In: CVPR. IEEE (2012)
36. Wang, H., Nie, F., Huang, H.: Robust and discriminative self-taught learning. In: International Conference on Machine Learning, pp. 298–306 (2013)
37. Wang, H., Nie, F., Huang, H.: Robust distance metric learning via simultaneous ℓ_1-norm minimization and maximization. In: ICML, pp. 1836–1844 (2014)
38. Wang, H., Nie, F., Huang, H., Yang, Y.: Learning frame relevance for video classification. In: Proceedings of the 19th ACM International Conference on Multimedia, pp. 1345–1348. ACM (2011)

39. Wang, H., Yan, L., Huang, H., Ding, C.: From protein sequence to protein function via multi-label linear discriminant analysis. IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) **14**(3), 503–513 (2017)
40. Wright, J., Ganesh, A., Rao, S., Peng, Y., Ma, Y.: Robust principal component analysis: exact recovery of corrupted. In: NIPS, p. 116 (2009)
41. Wright, S.J., Nocedal, J.: Numerical optimization. Springer Sci. **35**(67–68), 7 (1999)
42. Yuan, H., Paskov, I., Paskov, H., González, A.J., Leslie, C.S.: Multitask learning improves prediction of cancer drug sensitivity. Sci. Rep. **6**, 31619 (2016)