Drug-induced QT Prolongation Prediction Using Co-regularized Multi-view Learning

Jintao Zhang
Center for Bioinformatics
University of Kansas
Lawrence, KS 66047
Email: jtzhang@ku.edu

Jun Huan
Department of Electrical Engineering
and Computer Science
University of Kansas, Lawrence, KS 66045
Email: jhuan@ittc.ku.edu

***Abstract*—Drug-induced QT prolongation is a major life-threatening adverse drug effect. It is crucial to predict the QT prolongation effect as early as possible in drug development, however, data on drugs that induce QT prolongation are very limited and noisy. Multi-view learning (MVL) has been applied to many challenging machine learning and data mining problems, especially when complex data from diverse domains are involved and only limited labeled examples are available. Unlike existing MVL methods that use ℓ_2-norm co-regularization to obtain a smooth objective function, in this paper we proposed an ℓ_1-norm co-regularized MVL algorithm for predicting drug-induced QT prolongation effect and reformulate the ℓ_1-norm co-regularized objective function for deriving its gradient in the analytic form. ℓ_1-norm co-regularization enforces sparsity in the learned mapping functions and hence the results are expected to be more interpretable. Comprehensive experimental comparisons between our proposed method and previous MVL and single-view learning methods demonstrate that our method significantly outperforms those baseline methods.**

I. Introduction

Torsades de pointes (TdP) is a rare form of polymorphic ventricular tachycardia that exhibits distinct characteristics on the electrocardiogram (ECG), and is considered as a major life-threatening condition for patients as it can degenerate into ventricular fibrillation and sudden death [1]. In the past decade, the single most common cause of the withdrawal or restriction of the use of marketed drugs has been recognized as the prolongation of the QT interval associated with TdP [2]. Until 2004, at least nine drugs that were marketed in the U.S. or elsewhere have been removed from the market or had their availability restricted due to the risk of prolonging QT interval, such as Astemizole [3] and Grepafloxacin [4]. Therefore, it is crucial to discover and filter out those drug candidates with potential risk of QT prolongation and/or TdP as early as possible in the drug development pipeline to save development expenses and lives in clinical trials. Note that in this paper the terms QT prolongation and TdP risk are used exchangeably for ease of description.

TdP is often associated with a prolonged QT interval, which may be congenital or acquired. Drug-induced QT prolongation has triggered significant attention from pharmaceutical industry. Since 2005 the U.S. Food and Drug Administration (FDA) and European regulators have required that nearly all new molecular entities must be evaluated in a thorough QT study to determine its effects on the QT interval [5]. Although QT interval can be influenced by many factors, understanding of the underlying genetic mechanisms of QT prolongation have been significantly improved during the past decade. At present a few ion-channel related genes have been identified, and their gene products are found to profoundly influence the balance of ion-channel currents that determine the duration of the myocyte's action potential and thus the QT interval [6], including three potassium ion channels $I_{kr}(K_v 11.1$, hERG, KCNH2), $I_{k1}(K_{ir}2.1$, KCNJ2), and $I_{ks}(K_v 7.1$, KCNQ1), one sodium ion channel $I_{Na}(Na_v 1.5$, SCN5A), and one calcium ion channel $I_{CaL}(Ca_v 1.2$, CACNA1C) [1]. Presumably, compounds that inhibit one or more these involved ion channels are more likely to obstruct the normal cardiac ion conduction and thus prolong the QT interval. This key observation is crucial for identifying compounds with potential TdP risk.

There are various means for pharmaceutical companies to obtain QT prolongation information induced by their interested compounds, such as animal models and human clinical trials, whereas the FDA Adverse Event Reporting System is another important source for such information. However, data obtained from these sources are either expensive and time-consuming or noisy, while computationally predicting this rare but serious adverse drug effect (ADE) in humans remains highly challenging. With thorough literature search we identified 142 compounds [1], [7] that are possibly associated with QT prolongation and TdP with different risk levels, including 28 compounds labeled as "risk", 40 as "possible risk", and 74 as "conditional or congenital risk" (http://www.azcert.org). It is not surprising that most of these drug compounds have such uncertain labels, since it is non-trivial to determine the causation between drugs and TdP risk, especially when multiple drugs are simultaneously taken by patients. With such limited and noisy data, traditional machine learning methods cannot build accurate models for predicting the QT prolongation and TdP risks of potential drug candidates. On the other hand, additional knowledge on the mechanism of QT prolongation are available but unused.

A variety of applications have limited and noisy labeled examples such as QT prolongation prediction, where traditional single-view learning methods work poorly. Multi-view learning (MVL) has been applied to many challenging machine learning and data mining problems, especially when complex data from diverse domains are involved and only limited

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

labeled examples are available. The underlying assumption of multi-view learning [8]–[10] is that different views are conditional independent and relatively complementary to one another, so that combining knowledge from multiple views can afford performance gain. Among MVL algorithms in the co-regularization [8] framework, all previous work used ℓ_2-norm co-regularization for it is easy to optimize the smooth objective function. In this paper, we propose an ℓ_1-norm co-regularized MVL algorithm to boost prediction performance on the important bioinformatics problem: prediction of QT prolongation effect using the limited and noisy data available. Noticeably, although this MVL algorithm is developed for predicting the effect of QT prolongation, the method itself is domain independent and could be applied to a wide range of other real-world applications with limited and noisy multi-view data.

All previous MVL algorithms that use ℓ_2-norm co-regularization have a smooth objective function optimized by an alternate optimization approach [11], [12], i.e., in each iteration one view is optimized with the other views fixed until convergence of mapping functions in all views. However, we will demonstrate that our proposed ℓ_1-norm co-regularized MVL method is more advantageous after properly reformulating the objective function and computing its gradient in the analytic form. Moreover, We developed a simultaneous optimization approach that optimizes the mapping functions from all views simultaneously.

For predicting drug-induced QT prolongation using multi-view learning, the first view is easily directed to drug chemical structures, while there is a few options for the second view. One choice is the general binding profiles of the drug compounds to human protein. More specifically, we can use the binding information of compounds to those human proteins associated with QT prolongation, i.e. the five ion-channel proteins mentioned previously. However, it is challenging to obtain binding data between the compounds interested and the five QT prolongation-related ion-channel proteins mentioned above. When the binding activity between a compound and one of the ion channels is unknown or unavailable, we build protein-chemical interaction prediction models to predict it until the activity matrix is filled up completely. By learning these two-view data, ℓ_1-norm co-regularization enforces sparsity in the learned mapping functions and naturally carries the functionality of feature selection, and hence the prediction results are expected to be more accurate and interpretable. Comprehensive experimental comparisons are designed for demonstrating the advantages of our proposed ℓ_1-norm co-regularized MVL method over reference methods on the prediction of drug-induced QT prolongation.

II. Related Work

Many computational methods for facilitating the prediction of drug-induced QT prolongation and TdP risk have been developed and applied in various applications such as drug discovery and development. Yap *et al.* [13] used linear solvation energy relationships descriptors to measure the TdP-causing potential of compounds and built prediction models with support vector machines (SVMs). Their method simply ignored the different confidence levels of data and considered all "risk", "possible risk", and "conditional or congenital risk" compounds as positive samples, and the interpretability and meaning of their results were questionable. Recognizing the fact that the hERG protein is frequently associated with QT prolongation, Yao *et al.* [14] and Redfarn *et al.* [15] developed models using hERG inhibition activity and other relevant data to predict drug-induced QT prolongation, and investigated its relationships with various electrophysiological properties. The limitation is that hERG is not the only or necessary binding protein to induce this serious ADE. In addition, Champerous *et al.* [16] classified training examples into three categories according to their confidence levels, and developed a complex algorithm called TPDscreen^{TM} combined with a database from reference compounds with available clinical data to predict the risk of TdP and QT prolongation at the early stage of drug development.

Given the fact that the development of QT prolongation and TdP is associated with multiple genes [6], simulation studies of multiple ion channels blockade have been conducted by Mirams *et al.* [17]. They collected multiple ion channels (hERG, I_{Na}, and I_{CaL}) data on 31 drugs associated with various risk of TdP, and performed simulations with a series of mathematical models of cardiac cells to integrate information on multi-channel block, resulting in improved prediction of TdP risk. Ouillé *et al.* [1] picked five TdP-associated ion channels, including three potassium channels (I_{KR}, I_{K1}, and I_{KS}), one sodium (I_{Na}) and one calcium (I_{CaL}) channel, and elucidated the mechanisms of drug-induced TdP using ion channel blocking profiles. They studied the effects of known torsadogenic and non-torsadogenic compounds on these ion channels, investigated their functions in the generation of drug-induced TdP and QT prolongation, and identified that I_{Na} and I_{K1} play roles that are as important as I_{KR} in safety pharmacology. This study motivated us to use the blocking profiles of the five ion channels as the second view to improve our MVL models.

The rest of this paper is organized as follows: in Section II we introduce our new MVL methods in more detail. Data collection, feature extraction, and all experimental and comparison results are described in Section III. We then conclude our work in Section IV with insightful observations.

III. Methods

In our data set, each example has two view vectors. For labeled examples, the two view vectors are associated with the same label, while unlabeled examples doesn't have a label but can be helpful on improving model performance. In this section we discuss our proposed ℓ_1-norm co-regularized MVL algorithm in detail, including how we integrate information from all views and derive the problem formulation step by step. We also derive the analytic form of the gradients of our objective functions for simultaneous optimization, as compared to the traditional alternate optimization technique.

A. ℓ_1-norm Co-regularized Multi-view Learning

Suppose we have V views and for each view $v = 1, 2, ..., V$, use $X_v \in R^{N \times M_v}$ to denote the feature matrix with N

examples and M_v features on view v. Let x_i represents the concatenated row feature vector for example $i = 1, 2, ..., N$. The N examples can be partitioned into two subsets L and U across all views: a small number of labeled examples $\{(x_i, y_i)\}_{i \in L}$ in L, and a set of unlabeled example $\{x_i\}_{i \in U}$ in U, where $|L| \ll |U|$ and the label $y_i \in \{-1, 1\}$. Specifically, we use X_v^L and X_v^U to denote the feature matrix of labeled samples and unlabeled samples for view v. The basic idea underlying co-regularized multi-view learning is to learn one mapping function h^v on each view v so that these functions agree each other on an unlabeled example as much as possible while the error on the labeled examples is minimized. The final prediction function h is defined as:

$$h(x_i) = \frac{1}{V} \sum_{v=1}^{V} h^v(x_{i,v}), \qquad (1)$$

where $x_{i,v}$ represents the feature vector in the v view for example i. The view mapping functions h^1, h^2, ..., h^v are obtained by *minimizing* the following objective function over these view functions:

$$\begin{aligned} F(h^1, ..., h^v) &= \sum_{i \in L} L(y_i, h(x_i)) + \sum_{v=1}^{V} \lambda_v ||h^v||_p \\ &+ \mu \sum_{v \neq v_1}^{V} \|h^v(X_v) - h^{v_1}(X_{v_1})\|^2, \quad (2) \end{aligned}$$

where the first term is the loss function that penalizes the misclassification on labeled examples, the second term is to regularize the mapping function on each view with ℓ_p-norm ($1 \leq p \leq +\infty$) so that irregular solutions will be filtered out, and the final term is to penalize the disagreement among the prediction results from different view functions and to drive them to agree one another as much as possible. Here parameters $\{\lambda_j\}$ are the strength of ℓ_1-norm regularization terms in all views, and μ is the coupling parameter that regularizes the prediction disagreement using unlabeled data. By minimizing the sum of these three terms, co-regularization based MVL aims to identify a set of regular mapping functions that minimize misclassification rates on labeled examples and agree one and another on the predictions of unlabeled examples as much as possible.

For simplicity, without loss of generality we consider only two views and take least square loss on labeled examples and use ℓ_1-norm regularization on both views, resulting in the following simplified objective function,

$$\begin{aligned} F(h^1, h^2) &= \sum_{i \in L} \|y_i - h(x_i)\|_2^2 + \lambda_1 ||h^1||_1 + \lambda_2 ||h^2||_1 \\ &+ \frac{1}{2}\mu \sum_{i \in U} \|h^1(x_i^1) - h^2(x_i^2)\|_2^2, \qquad (3) \end{aligned}$$

By applying a linear mapping function W_v on view $v = 1, 2$, $h^v(X_v) = X_v W_v$, where W_v is a $M_v \times 1$ column coefficient vector for view v, we have a metricized objective function as follows:

$$\begin{aligned} F(W_1, W_2) &= \|Y - \frac{1}{2}(X_1^L W_1 + X_2^L W_2)\|_2^2 + \lambda_1 ||W_1||_1 \\ &+ \lambda_2 ||W_2||_1 + \frac{1}{2}\mu \|X_1^U W_1 - X_2^U W_2\|_2^2, (4) \end{aligned}$$

If ℓ_2-norm co-regularization is used in Equation (4), this objective function is convex and smooth, and is usually optimized using the alternate optimization approach [11], [12], i.e., alternately optimizing one view with the other view fixed until convergence. When we use ℓ_1-norm co-regularization as in Eqn.(4), the objective function becomes convex but not smooth. However, by reformulating this function properly we can optimize both views simultaneously. Let [.] denote the horizontal concatenation of two matrices $Z_1 \in R^{n \times p}$ and $Z_2 \in R^{n \times q}$, and then $[Z_1 \ Z_2] \in R^{n \times (p+q)}$. We concatenate column vectors W_1, W_2 and matrices X_1, X_2 as $W = [W_1^T \ W_2^T]^T$, $X_L = [X_1 \ X_2]_{x \in L}$, and $X_U = [X_1 \ -X_2]_{x \in U}$. The objective function F in Eqn.(4) can be rewritten as

$$\begin{aligned} F(W) &= \|Y - \frac{1}{2}X_L W\|^2 + \lambda^T \|W\|_1 + \frac{1}{2}\|X_U W\|^2, \\ &= (Y - \frac{1}{2}X_L W)^T (Y - \frac{1}{2}X_L W) + \lambda^T \|W\|_1 \\ &+ \frac{1}{2}\mu (X_U W)^T X_U W, \qquad (5) \end{aligned}$$

where (λ) is a $(M_1 + M_2) \times 1$ constant vector with the first M_1 elements as λ_1 and the remaining M_2 elements as λ_2. It is straightforward that we can derive the analytic form of the gradient of the function in Equation (5) as follows:

$$\begin{aligned} \frac{\partial F(W)}{\partial W} &= -X_L^T (Y - \frac{1}{2}X_L W) + \mu X_U^T X_U W \\ &+ \lambda * \text{sign}(W), \qquad (6) \end{aligned}$$

We then use the LBFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) method [18], [19] to minimize the objective function in Equation (5). With the known analytic form of the gradient as in Equation (6), the optimization process can be very efficient.

B. ℓ_2-norm Co-regularized Multi-view Learning

The ℓ_2-norm co-regularized MVL method has been discussed in many previous work [8], [9]. The objective function of this MVL method is as follows,

$$\begin{aligned} F(W_1, W_2) &= \|Y - \frac{1}{2}(X_1^L W_1 + X_2^L W_2)\|_2^2 + \lambda_1 ||W_1||_2 \\ &+ \lambda_2 ||W_2||_2 + \frac{1}{2}\mu \|X_1^U W_1 - X_2^U W_2\|_2^2, (7) \end{aligned}$$

Some previous works use an alternate optimization method to identify optimal W_1 and W_2. The alternate process can be very time-consuming, however we can apply some reformulation tricks and implement it in a more efficient matter. Similar to ℓ_1-norm co-regularization, the gradient of this objective function can be written as,

$$\begin{aligned} \frac{\partial F(W)}{\partial W} &= -X_L^T (Y - \frac{1}{2}X_L W) + \mu X_U^T X_U W \\ &+ \frac{\lambda_1}{\|W_{11}\|_2} W_{11} + \frac{\lambda_2}{\|W_{22}\|_2} W_{22}, \qquad (8) \end{aligned}$$

TABLE I
SUMMARY OF DATA AND RESULTS FOR THE SECOND VIEW GENERATION. "IC" - ION-CHANNEL PROTEINS.

IC	AID	Active	Inactive	# Used	Accuracy
I_{Na}	/	8	17	All	0.788
I_{CaL}	/	53	24	All	0.750
I_{kr}	376	250	1703	200	0.789
I_{ks}	2642	3878	301,738	200	0.778
I_{k1}	2032	926	1,338	200	0.739

where variables W, Y, X_L, X_U, and U have the same meaning as in the previous subsection. W_{11} and W_{22} are vectors with the same length of W, and are obtained by appending 0's at the end of the original vector W_1 and at the beginning of W_2, respectively. With this gradient we can optimize the objective function in Equation (7) similarly as for our ℓ_1-norm co-regularized MVL method.

IV. RESULTS

In this section, we began our discussion with an introduction to how we collected, cleaned, and represented the TdP-related drugs and those compounds with known binding activity to available human proteins, including the five TdP-associated ion-channel proteins. For performance comparison, three reference methods are selected as comparison baselines. SVMs [20] is a widely used single-view supervised learning algorithm with excellent performance. The LIBSVM [21] implementation and RBF kernels were used in all experiments. As a generalized linear method, ℓ_1-regularized logistic regression (LLR) [22] is another excellent single-view supervised learning method can construct a model that estimates probabilities, e.g. for medical diagnosis and credit scoring. In this paper, we used the algorithm proposed by Boyd *et al.* [22] to run our experiments of the LLR method. The third reference method is the ℓ_2-norm co-regularized MVL method, which has been discussed in many previous works [8], [9] and introduced in the METHODS section.

A. Data Set and Feature Extraction

We collected the compound set of drug-induced QT prolongation from the home page of University of Arizona CERT (http://www.azcert.org/) [7] and the work by Ouillé *et. al* [1]. We merged these two data sets by removing redundant and/or inconsistent data points, resulting in a set of 28 drug compounds labeled as "TdP risk", 40 labeled as "possible TdP risk", and 74 with other less reliable evidence of QT prolongation. In our experiments we treated the 28 drugs with "TdP risk" labels as positive examples, the 40 drugs with "possible TdP risk" labels as positive or unlabeled examples, and the remaining 74 compounds as unlabeled examples whose labeled to be determined by our MVL models. For negative examples, we extracted all approved drugs from the DrugBank database [23], excluded all drug compounds that are associated with QT prolongation and/or TdP risk with even very weak evidence, and obtained a set of 1,221 drug compounds as putative negative examples.

We used molecular signature descriptors [24] to extract features from chemical structures. In a molecular signature vector, each component is the number of occurrences of a particular atomic signature in the molecule. An atomic signature is a canonical representation of the substructure surrounding a particular atom, including all atoms and bonds up to a predefined number (called the signature height) of consecutive bonds from the given atom. We set the signature height at 2 for all compounds, filtered out all atomic signatures with frequency less than 4%, and finally obtained about 100 features for each compound. We used drug signature descriptors as the first view of our data set.

In our experimental study, we first randomly selected an equal number of negative and positive examples and combined them with unlabeled data to make a balanced subset, from which we randomly selected 20% positive and 20% negative examples for testing. We used the remaining 80% labeled and all unlabeled examples for training with 10-fold cross validation to select the best model parameters. For each experimental setting, we repeated the experiment for 50 times using random sampling, and reported the mean and standard deviation of testing accuracy, precision, and recall, which are defined as (TP+TN)/S, TP/(TP+TN), and TP/(TP+FN), respectively, where TP is true positive, TN is true negative, FN is false negative, and S is the total number of testing examples.

B. Generating the second view

In the MVL framework, each data example has multiple feature vectors, one from each view. Our first view is taken from compound structures also using molecular signature descriptors [24]. We set signature distance at 2 and frequency threshold at 4%, and ended up with a vector of ∼90 non-negative features for each compound. We first used the binding profiles of all drug compounds to the five ion-channel proteins as the second view. For the binding activity that are unknown or unavailable, we have built protein-chemical interaction prediction models to find appropriate decisions, resulting in a vector of five binary elements, in which a bit is set to 1 if a drug compound inhibits an ion channel, and 0 otherwise. To this end, we manually searched for binding compounds of the five picked ion channels from the PubChem database. For the potassium ion channels I_{KR}, I_{K1}, and I_{KS}, we identified multiple target-based bioassays that use them as target proteins, and then selected bioassays with AID 376, 2032 and 2642 for each of them, respectively. Each assay provides hundreds of active compounds and a large number of inactive compounds. For the calcium ion channel I_{CaL} and the sodium ion channel I_{Na}, we identified multiple related target-based assays with few tested compounds. We decided to use those tested compounds with activity less than 1μM as active and the rest as inactive, and obtained 8 (53) weakly active and 17 (24) inactive compounds for the ion channel I_{Na} (I_{CaL}).

We randomly selected 100 active and 100 inactive compounds for the potassium ion channels as a balanced data subset and used all active and inactive compounds identified for the sodium and calcium channel. The model construction and selection process is as follows: first 20% active and 20% inactive compounds were randomly selected as the testing set and the rest 80% as the training set, and 10-fold cross validation were applied to the training set with support vector

TABLE II
PREDICTING ACCURACY USING ION-CHANNEL BINDING PROFILES. STANDARD DEVIATIONS INCLUDED IN PARENTHESIS.

		$\lvert X_L \rvert$	= 28			$\lvert X_L \rvert$	= 68	
Method	$\lvert X_u \rvert$	Accuracy	Precision	Recall	$\lvert X_u \rvert$	Accuracy	Precision	Recall
SVM	/	0.727(0.060)	0.779(0.096)	0.725(0.105)	/	0.687(0.049)	0.692(0.058)	0.696(0.039)
LLR	/	0.712(0.078)	0.779(0.105)	0.658(0.133)	/	0.676(0.074)	0.701(0.090)	0.657(0.084)
ℓ_1-MVL	114	**0.771(0.070)**	**0.802(0.085)**	**0.825(0.096)**	74	**0.716(0.038)**	0.707(0.051)	**0.723(0.054)**
ℓ_2-MVL	114	0.748(0.076)	0.739(0.077)	0.817(0.091)	74	0.693(0.062)	0.706(0.061)	0.686(0.074)
ℓ_2-MVL-ALT	114	0.744(0.075)	0.736(0.080)	0.805(0.076)	74	0.696(0.066)	0.697(0.059)	0.691(0.074)

machines (SVMs) and RBF kernels to build one model for each ion channel using its own binding compounds. In 10-fold cross validation, the training data was randomly split into 10 equal disjoint subsets. Nine subsets were used for training a model, and the rest subset was used as a validation set for validating the model. The cross-validation process repeated for 10 times with each subset used exactly once as the validation set. Each experiment was repeated for ten times, and a compound was considered to be active to an ion channel if it was predicted as active for six or more times. Eventually our models have an average accuracy of about 73-78%, as shown in Table I.

C. Experimental Results

In this section, we compared our proposed ℓ_1-norm co-regularized MVL algorithm (ℓ_1-MVL) with SVMs, ℓ_1-regularized logistic regression (LLR), and the ℓ_2-norm co-regularized MVL method using both our proposed optimization (ℓ_2-MVL) and alternate optimization (ℓ_2-MVL-ALT) techniques, and evaluated their performance using testing accuracy, precision, and recall. Note that our results were obtained from 20 repeated experiments, so the performance differences between different methods should be considered significant. We tested these five methods at various experimental settings to investigate the contribution and influence of the related factors on prediction performance.

First we use the ion-channel binding profiles as the second view, and the data set consists of 28 positive, 40 possible positive, 74 unlabeled and 1,221 putative negative examples. If we used only 28 positive examples and treated the 40 possible positive examples as unlabeled, we then have 114 unlabeled examples. The results of the five methods were summarized in the first five rows in Table II. We found that our ℓ_1-MVL method significantly outperformed not only the SVL methods (SVMs and LLR) for 5-6%, but also the ℓ_2-MVL method for about 3.3% on prediction accuracy. For precision, the ℓ_1-MVL method was slightly better than SVMs and LLR, and much better than ℓ_2-MVL methods. All MVL methods have much better recall than SVMs and LLR. All these performance metrics consistently reveal that our proposed ℓ_1-MVL method performs better than the reference methods. Moreover, to investigate the contribution of the unlabeled examples to prediction performance, we replaced the matrix X_U with a zero matrix of the same dimensionality, and performed similar experiments described above. We listed the results in the last three rows in Table II. We observed apparent decrease of the ℓ_1-MVL method on accuracy and recall, while the precision results are still comparable or slightly better than previous experiments. We can approximately estimate the contribution of unlabeled examples in the the ℓ_1-MVL method is about 1.1%, which is non-trivial considering that the total accuracy improvement of our MVL method is only 5-6% compared to SVL methods. Although the ion-channel binding profiles has very small dimensionality as the second view, its contribution to the performance improvement is significant by comparing the results of the MVL methods with the SVL methods.

If we gave them more confidence on the 40 drug compounds labeled as "possible TdP risk"and used them as positive examples, we would have more labeled and less unlabeled examples. With 68 positive examples, we did similar experiments to investigate the influence of the number of labeled examples on prediction performance. All results with the same experimental settings as above are summarized in right columns in Table II. By using those "possible positive examples" as positive examples, we observed significant performance decrease on all methods. Although the ℓ_1-norm co-regularized MVL method is still better than the baseline methods, the performance advantage shrank to 1-2%. By adding unreliable positive examples to our data set, we introduced many false positives and thus compromised the prediction accuracy. Moreover, the precision of the ℓ_1-MVL method decreased about 10%, which more clearly demonstrates that the data set contains much more false positives than before, according to the definition of precision. We also found that the precision of ℓ_2-MVL method dropped only about 3%. The observation showed that our proposed ℓ_1-MVL method is more sensitive to false positive examples. Finally, it is easy to notice that the performance of the MVL methods fluctuated very slightly when we have different numbers of unlabeled examples, apparently because the contribution of unlabeled examples have been offset by introducing more false positive examples.

D. Contribution of the Second View

In this section we discussed the contribution of the ion-channel protein binding profiles of the drug compounds as the second view to final models, since this view significantly improved the performance of our ℓ_1-norm co-regularized MVL method, comparing with the single-view learning methods such as SVM and LLR. In the final models, we extracted the model coefficients of these five binary features from the second view, and calculated the mean and standard deviation of these coefficients under the four different experimental setting previously described. According to the fourth chapter in [25], the mean to standard deviation ratio, called Z-score, of a feature's model coefficient measures the significance of this feature's contribution to the final models. From Table III we can find that the ion channel I_{K1} plays a relatively more important role since its Z-score ∈ [0.92,1.33] is much greater

TABLE III
Z-SCORES OF MODEL COEFFICIENTS FROM ℓ_1-REGULARIZED LOGISTIC REGRESSION FOR THE FIVE ION-CHANNEL PROTEINS.

$\lvert X_L\rvert$	$\lvert X_u\rvert$	I_{Na}	I_{CaL}	I_{KR}	I_{KS}	I_{K1}
28	114	0.308	0.061	0.300	0.082	1.327
28	0	-0.134	-0.077	0.042	-0.024	1.916
68	74	0.337	0.039	0.615	-0.090	0.923
68	0	0.434	-0.020	0.454	-0.097	0.968

than other ion channels. This finding also partly matches the observation in [1], in which ion channel I_{Na} and I_{K1} are found to play a key role in the generation of drug-induced QT prolongation and TdP. In our models the importance of the ion channel I_{Na} was not identified, mainly because we have very few training examples (8 weak positives and 17 negatives) available for this ion channel in the data set. In addition, the Z-scores of I_{CaL} and I_{KS} are very close to zero, and can be reasonably considered as the evidence of the sparsity of our final models, i.e., two our of the five ion-channel feature was considered no significant relevance, and we can set the corresponding coefficients as zero.

V. Discussion

In this paper we aim to construct accurate prediction models with very limited and noisy data that are available, and propose a novel ℓ_1-norm co-regularized multi-view learning algorithm on predicting drug-induced QT prolongation. We conduct experiments under various settings to investigate the contribution and influence of the number of unlabeled examples and labeled examples. Experimental results show that our ℓ_1-MVL method outperforms not only two excellent SVL methods (SVMs and LLR), but also the widely used ℓ_2-norm co-regularized MVL method with the measurement of testing accuracy, precision, and recall.

To handle the challenge of limited and noisy labeled examples, we extract the second view for all TdP related drug compounds from critical related works, and pick five ion-channel proteins that have been found highly relevant to the QT prolongation effect. The contribution of the second view has been clearly demonstrated by our comprehensive experiments and significant higher prediction accuracy in both the ℓ_1-norm and ℓ_2-norm co-regularization framework. From the coefficients in the final models, we analyzed the significance of the contribution of each ion channel to QT prolongation, and our finding partly matched previous work [1].

Acknowledgment

This work has been supported by the KU Specialized Chemistry Center (NIH U54 HG005031) and NIH grant R01 GM868665.

References

[1] A. Ouillé et al., "Ion channel blocking profile of compounds with reported torsadogenic effects: what can be learned?" *Br. J. Pharmacol.*, March 2011.

[2] K. E. Lasser et al., "Timing of new black box warnings and withdrawals for prescription medications." *J. Am. Med. Asso.*, vol. 287, no. 17, pp. 2215–2220, 2002.

[3] F. de Abajo and L. Rodríguez, "Risk of ventricular arrhythmias associated with nonsedating antihistamine drugs." *Br. J. Clin. Pharmacol.*, vol. 47, no. 3, pp. 307–313, 1999.

[4] P. Ball, "Quinolone-induced QT interval prolongation: a not-so-unexpected class effect." *J. Antimicrob. Chemother.*, vol. 45, no. 5, pp. 557–559, 2000.

[5] FDA, "E14 clinical evaluation of QT/QTc interval prolongation and proarrhythmic potential for non-antiarrhythmic drugs," October 2005.

[6] A. J. Moss, "Drug-induced QT prolongation: an update," *Ann. Noninvasive Electrocardiol.*, vol. 11, no. 1, pp. 1–2, 2006.

[7] R. L. Woosley, "Drugs that prolong the QT interval and/or induce torsades de pointes ventricular arrhythmia," University of Arizona CERT, 2003.

[8] V. Sindhwani and P. Niyogi, "A co-regularized approach to semi-supervised learning with multiple views," in *Proceedings of the ICML Workshop on Learning with Multiple Views*, 2005.

[9] V. Sindhwani and D. S. Rosenberg, "An rkhs for multi-view learning and manifold co-regularization," in *Proceedings of ICML'08*, 2008, pp. 976–983.

[10] M. Culp, G. Michailidis, and K. Johnson, "On multi-view learning with additive models," *Ann. Applied Stat.*, vol. 3, no. 1, pp. 292–318, 2009.

[11] S. Yu et al., "Bayesian co-training." in *Proceedings of NIPS'07*, 2007.

[12] B. Krishnapuram et al., "On semi-supervised classification." 2004.

[13] C. Yap et al., "Prediction of torsade-causing potential of drugs by Support Vector Machine approach," *Toxicol. Sci.*, vol. 79, no. 1, pp. 170–177, 2004.

[14] X. Yao et al., "Predicting QT prolongation in humans during early drug development using hERG inhibition and an anaesthetized guinea-pig model," *Br. J. Pharmacol.*, vol. 154, no. 7, pp. 1446–1456, 2008.

[15] W. Redfern et al., "Relationships between preclinical cardiac electrophysiology, clinical QT interval prolongation and torsade de pointes for a broad range of drugs: evidence for a provisional safety margin in drug development." *Cardiovasc. Res.*, vol. 58, no. 1, pp. 32–45, 2003.

[16] P. Champeroux et al., "Prediction of the risk of torsade de pointes using the model of isolated canine Purkinje fibres," *Br. J. Pharmacol.*, vol. 144, no. 3, pp. 376–385, 2005.

[17] G. R. Mirams et al., "Simulation of multiple ion channel block provides improved early prediction of compounds' clinical torsadogenic risk," *Cardiovasc. Res.*, March 2011.

[18] H. Matthies and G. Strang, "The solution of non linear finite element equations." *Intl. J. Num. Methods in Engineering*, vol. 14, pp. 1613–1626, 1979.

[19] J. Nocedal, "Updating quasi-Newton matrices with limited storage." *Mathematics of Computation*, vol. 35, pp. 773–782, 2004.

[20] C. J. Burges, "A tutorial on support vector machines for pattern recognition." *Data Mining and Knowledge Discovery*, vol. 2, no. 4, pp. 121–167, 1998.

[21] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[22] K. Koh, S.-J. Kim, and S. Boyd, "An interior-point method for large-scale ℓ_1-regularized logistic regression." *J. Mach. Lear. Res.*, vol. 8, pp. 1519–1555, 2007.

[23] D. S. Wishart et al., "DrugBank: a knowledgebase for drugs, drug actions and drug targets." *Nucl. Acids Res.*, vol. 36, pp. D901–906, 2008.

[24] J.-L. Faulon, M. Collins, and R. Carr, "The signature molecular descriptor. 4. canonizing molecules using extended valence sequences." *J. Chem. Inf. Model.*, vol. 44, no. 2, pp. 427–436, 2004.

[25] T. Hastie, R. Tibshirani, and J. Friedman, *Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition*. Springer, 2009.