Optimizing Multivariate Performance Measures from Multi-View Data

Jim Jing-Yan Wang[1], **Ivor Wai-Hung Tsang**[2], **Xin Gao**[1] *
[1] King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE), Thuwal, 23955-6900, Saudi Arabia
[2]Center for Quantum Computation and Intelligent Systems, University of Technology Sydney, Australia
jimjywang@gmail.com, ivor.tsang@uts.edu.au, xin.gao@kaust.edu.sa

Abstract

To date, many machine learning applications have multiple views of features, and different applications require specific multivariate performance measures, such as the F-score for retrieval. However, existing multivariate performance measure optimization methods are limited to single-view data, while traditional multi-view learning methods cannot optimize multivariate performance measures directly. To fill this gap, in this paper, we propose the problem of optimizing multivariate performance measures from multi-view data, and an effective method to solve it. We propose to learn linear discriminant functions for different views, and combine them to construct an overall multivariate mapping function for multi-view data. To learn the parameters of the linear discriminant functions of different views to optimize a given multivariate performance measure, we formulate an optimization problem. In this problem, we propose to minimize the complexity of the linear discriminant function of each view, promote the consistency of the responses of different views over the same data points, and minimize the upper boundary of the corresponding loss of a given multivariate performance measure. To optimize this problem, we develop an iterative cutting-plane algorithm. Experiments on four benchmark data sets show that it not only outperforms traditional single-view based multivariate performance optimization methods, but also achieves better results than ordinary multi-view learning methods.

Introduction

In different machine learning applications, different multivariate performance measures are used for the purpose of performance evaluation. For example, in problems of text classification, F1-score and precision/recall breakeven point (PRBEP) are used to compare true class labels against predicted class labels of a given test text set. In image retrieval problems, the area under the receiver operating characteristic curve (AUROC) is used to evaluate the performance of a retrieval system. However, a classifier trained by optimizing a common loss function over a training set, such as a hinge loss or a logistic loss, cannot guarantee that an optimal multivariate performance measure can be obtained over a test set, such as an F-score.

To solve this problem, recently, the problem of multivariate performance measure optimization was proposed to learn a classifier by directly optimizing a desired multivariate performance measure, instead of a common loss, over a training set (Joachims 2005). Since then, a number of state-of-the-art methods were proposed. For example, Joachims (2005) proposed a support vector machine (SVM)-based method, SVM^{perf}, to solve this problem, by training a multivariate SVM in polynomial time for multivariate performance measures by using a cutting-plane method (Kelley 1960). Zhang, Saha, and Vishwanathan (2012) proposed to improve the convergence rate of the cutting-plane method, and developed a novel smoothing strategy for the problem of multivariate performance measure optimization, by using the Nesterov's accelerated gradient method. Li, Tsang, and Zhou (2013) proposed a two-step approach to optimize multivariate performance measures, by first training a classifier with existing learning methods, and then adapting it to optimize a specific performance measure. Mao and Tsang (2013) proposed a novel feature selection method to optimize multivariate performance measures, by formulating the problem for high-dimensional data and employing a two-layer cutting-plane algorithm to solve it. Yang et al. (2015) proposed a novel multivariate performance optimization method based on sparse coding and hyper-predictor learning, by presenting the tuple of data points to a tuple of sparse codes and applying a linear function to compare a sparse code against a given candidate class label. Parambath, Usunier, and Grandvalet (2014) proposed to reduce the optimization of F-measure to a series of cost-sensitive classification problems. Koyejo et al. (2014) proposed to learn optimal classifiers for a family of performance measures, as the sign of the thresholded conditional probability of the positive class, using performance measure-dependent thresholds. Narasimhan, Kar, and Jain (2015) proposed an adaptive linearization approach for two families of performance measures which can be expressed as functions of true positive/negative rates. Recently, Wang and Gao (2015) proposed to optimize multivariate performance measures from partially labeled data tuple, by learning a classifier and completing the label tuple simultaneously.

Up to now, all these multivariate performance measure optimization methods are limited to learning from data with

*All correspondence should be addressed to Xin Gao. E-mail: xin.gao@kaust.edu.sa. Tel: +966-12-808-0323.

a single view. For example, when these methods are applied to image classification problems, only the features of the visual view are used as inputs. However, there is another type of view, i.e., user tags of images. Using only the visual view, one may not be able to present the data comprehensively, while leveraging data of multiple views can better present the data, due to the complementarity among different views. Learning from multiple views of data is referred as multi-view learning, and many methods have been proposed for this problem (Xu, Tao, and Xu 2013; Sun 2013). For example, Farquhar et al. (2005) proposed to learn an SVM classifier for each view, and imposed that the SVM classification responses of two different views should be consistent over the same data point. The two-view SVM was learned by minimizing the dissimilarity between the two-view responses measured by an ℓ_1-norm distance and optimizing a hinge loss. White et al. (2012) assumed that multi-view presentations of data have a shared latent representation, and proposed a multi-view subspace learning method to enforce conditional independence of different views while reducing dimensionality and recovering the optimal data reconstruction. Li et al. (2012) and Xu et al. (2015) proposed the co-labeling algorithm for the multi-view learning problems with uncertain labels, by learning multi-kernel classifiers and the optimal training labels simultaneously. Xu, Tao, and Xu (2015) proposed a multi-view intact space learning algorithm to discover a latent intact representation of the data by integrating the encoded complementary information in multiple views, by using the Cauchy loss which is robust to outliers. Mao et al. (2015) proposed a novel multi-kernel learning method to handle multi view data with partial correspondence and missing labels. However, to best of our knowledge, all the existing multi-view learning methods optimize some common loss functions in the training process, such as a hinge loss function, but ignore the multivariate performance measure used in the test process of a specific application. For example, in the two-view SVM method, the SVM classifiers are learned by optimizing the hinge loss (Farquhar et al. 2005), and in the co-labeling algorithm, the multi-kernel classifiers are learned by optimizing the squared hinge loss (Li et al. 2012). As a result, the learned classifier is not an optimal classifier for the desired multivariate performance measure.

In many real-world applications, it is necessary to learn classifiers from multi-view data to optimize multivariate performance measures. For example, in the multi-class image classification problem on the PASCAL VOC'07 data set, there are 20 classes, and researchers usually use a one-vs-all strategy for this multi-class problem. When images of one class are treated as positive data points, the images of the remaining nineteen classes are considered as negative data points, and the number of positive and negative data points are highly imbalanced, making the problem an imbalanced classification problem. In this case, we prefer to use some multivariate performance measures to evaluate the classification performance, such as the F1-score and AUROC. However, in many image data sets, the images have features from multiple views, e.g., images of PASCAL VOC'07 data set have two views, which are the visual view and the user tag view. The challenge of this problem lies on the imbalanced, noisy, and inconsistent multi-view data. To overcome this challenge, we propose to optimize the multivariate performance measures directly from multi-view data. The motivation is based on the observation that multivariate performance optimization methods are more sensitive to imbalanced and noisy data. Moreover, to solve the problem of inconsistency of multi-view data, we propose a regularizer, which uses consistency among different views to handle the noise. The necessity of imposing the multi-view consistency to performance measure optimization lies on the observation that, if the predictions of differen views over the same data point can lead to an optimal performance measure, these predictions should be consistent.

The contributions of this paper are of two folds:

1. We propose the problem of optimizing multivariate performance measures from multi-view data. Given a tuple of data points, each data point is presented by multiple views, the problem is to learn a multivariate mapping function to map them to a tuple of class labels, so that a desired multivariate performance measure can be optimized.
2. We proposed to learn linear discriminant functions for different views and combine them to construct an overall multivariate mapping function to predict the class label tuple for a tuple of data points. To learn the linear discriminant function parameters of different views, we formulate a constrained minimization problem. In this problem, we proposed to minimize the complexity of each linear discriminant function parameters by minimizing its squared ℓ_2-norm. We introduce a regularization to promote consistency among different views, by minimizing the squared ℓ_2-norm distances between responses of linear discriminant functions of each pair of different views over each data point. We also minimize the loss function corresponding to a specific multivariate performance measure. The minimization problem is optimized by a cutting-plane method in an alterative algorithm.

Proposed Method

Problem formulation

Assume we have a training data set of m views of n data points, presented as m data tuples, $\overline{\mathbf{x}}^j|_{j=1}^m$, where $\overline{\mathbf{x}}^j = (\mathbf{x}_1^j, \cdots, \mathbf{x}_n^j)$ is the data tuple of the j-th view of the n data points, and $\mathbf{x}_i^j \in \mathbb{R}^{d_j}$ is the d_j-dimensional feature vector of the j-th view of the i-th data point. The class label tuple of the n data points are given as $\overline{y} = (y_1, \cdots, y_n)$, where $y_i \in \{+1, -1\}$ is the binary class label of the i-th data point. We propose to learn a multivariate mapping function to predict a class label tuple for the multi-view data tuple $\overline{\mathbf{x}}^j|_{j=1}^m$, which is denote as $\overline{y}^* = (y_1^*, \cdots, y_n^*)$, where $y_i^* \in \{+1, -1\}$ is the predicted label of the i-th data point. To measure the performance of the prediction, we use a multivariate loss function $\Delta(\overline{y}^*, \overline{y})$, which corresponds to a desired multivariate performance measure, to compare the predicted label tuple, $\overline{y}^*$, with the true label tuple, $\overline{y}$. The multivariate performance measure optimization is to learn an op-

timal multivariate mapping function so that the multivariate loss function $\Delta(\overline{y}^*, \overline{y})$ can be minimized.

To implement this multivariate mapping function, we use a linear discriminant function $f_j(\overline{\mathbf{x}}^j, \overline{y}')$ to match the data tuple of the j-th view, $\overline{\mathbf{x}}^j$, with a candidate class label tuple $\overline{y}' = (y'_1, \cdots, y'_n)$,

$$f_j(\overline{\mathbf{x}}^j, \overline{y}') = \mathbf{w}_j^\top \Psi(\overline{\mathbf{x}}^j, \overline{y}'), \tag{1}$$

where$\Psi(\overline{\mathbf{x}}^j, \overline{y}) = \sum_{i=1}^n y'_i \mathbf{x}_i^j \in \mathbb{R}^{d_j}$ is a function which returns a d_j-dimensional vector to describe the match between $\overline{\mathbf{x}}^j$ and $\overline{y}'$, and $\mathbf{w}_j \in \mathbb{R}^{d_j}$ is a parameter vector for the linear discriminant function of the j-th view. We obtain a multi-view discriminant function by a linear combination of the discriminant functions of all the m views, and use it to match the multi-view data tuple against a candidate label tuple. The label tuple which maximizes the multi-view discriminant function is obtained as the predicted optimal label tuple,

$$\overline{y}^* = \arg\max_{\overline{y}' \in \mathcal{Y}} \left\{ \sum_{j=1}^m f_j(\overline{\mathbf{x}}^j, \overline{y}') = \sum_{j=1}^m \mathbf{w}_j^\top \Psi(\overline{\mathbf{x}}^j, \overline{y}') \right\}, \tag{2}$$

where $\mathcal{Y} = \{+1, -1\}^n$ is the set of all admissible label tuples. To simplify the denotation, we define an extended parameter vector, $\mathbf{w} = [\mathbf{w}_1{}^\top, \cdots, \mathbf{w}_m{}^\top]^\top \in \mathbb{R}^d$, by concatenating the m parameter vectors of the m views to one single vector, where $d = \sum_{j=1}^m d_j$. For the j-th view, the parameter vector $\mathbf{w}_j$ can be recovered by a view indicator matrix Θ_j,

$$\mathbf{w}_j = \Theta_j \mathbf{w}, \tag{3}$$

where $\Theta_j \in \{1, 0\}^{d_j \times d}$ is a $d_j \times d$ matrix of ones and zeros, and its (k, k')-th element $[\Theta_j]_{kk'} = 1$ if the k'-th element of $\mathbf{w}$ is the k-th element of $\mathbf{w}_j$, and 0 otherwise. To learn the parameter vectors $\mathbf{w}_j|_{j=1}^m$ for the m views, i.e., $\mathbf{w}$, we consider the following three problems.

1. **Reducing the complexity of each linear discriminative function**: To prevent over-fitting, we propose to reduce the complexity of the linear discriminative function of each view by minimizing the squared ℓ_2-norm of its parameter,

$$\min_{\mathbf{w}} \left\{ \frac{1}{2} \sum_{j=1}^m \|\mathbf{w}_j\|_2^2 = \frac{1}{2} \sum_{j=1}^m \mathbf{w}^\top \Theta_j^\top \Theta_j \mathbf{w} = \frac{1}{2} \mathbf{w}^\top \left(\sum_{j=1}^m \Theta_j^\top \Theta_j \right) \mathbf{w} = \frac{1}{2} \mathbf{w}^\top \mathbf{w} \right\}. \tag{4}$$

   Note here that $\sum_{j=1}^m \Theta_j^\top \Theta_j = I_{d \times d}$.

2. **Promoting consistency among different views**: Given the two views, $\mathbf{x}_i^j$ and $\mathbf{x}_i^{j'}$, of the i-th data point, their responses of the discriminative functions are $\mathbf{w}_j^\top \mathbf{x}_i^j$ and $\mathbf{w}_{j'}^\top \mathbf{x}_i^{j'}$, respectively. If the predictions of both views can lead to the same optimal desired multivariate performance, their responses should be consistent. This observation is inspired by multi-view consistency assumption proposed by White et al. (2012). Based on this observation, to promote the consistency of different views, we proposed to minimize the squared ℓ_2-norm distances of responses of the linear discriminative functions of each pair of views over the same data point,

$$\min_{\mathbf{w}_j|_{j=1}^m} \left\{ \frac{1}{2} \sum_{i=1}^n \left(\sum_{j,j':j<j'} \|\mathbf{w}_j^\top \mathbf{x}_i^j - \mathbf{w}_{j'}^\top \mathbf{x}_i^{j'}\|_2^2 \right) \right\}, \tag{5}$$

   By substituting (3) to (5), we have

$$\begin{aligned} \min_{\mathbf{w}} \Bigg\{ & \frac{1}{2} \sum_{i=1}^n \left(\sum_{j,j':j<j'} \|\mathbf{w}^\top \Theta_j^\top \mathbf{x}_i^j - \mathbf{w}^\top \Theta_{j'}^\top \mathbf{x}_i^{j'}\|_2^2 \right) \\ & = \frac{1}{2} \sum_{i=1}^n \left(\sum_{j,j':j<j'} \mathbf{w}^\top \left(\Theta_j^\top \mathbf{x}_i^j - \Theta_{j'}^\top \mathbf{x}_i^{j'} \right) \left(\Theta_j^\top \mathbf{x}_i^j - \Theta_{j'}^\top \mathbf{x}_i^{j'} \right)^\top \mathbf{w} \right) = \frac{1}{2} \mathbf{w}^\top \Lambda \mathbf{w} \Bigg\}, \end{aligned} \tag{6}$$

   where

$$\Lambda = \sum_{i=1}^n \sum_{j,j':j<j'} \left(\Theta_j^\top \mathbf{x}_i^j - \Theta_{j'}^\top \mathbf{x}_i^{j'} \right) \left(\Theta_j^\top \mathbf{x}_i^j - \Theta_{j'}^\top \mathbf{x}_i^{j'} \right)^\top \in \mathbb{R}^{d \times d}. \tag{7}$$

   In this way, we formulate the problem of view consistency as a minimization problem of a quadratic function of $\mathbf{w}$.

3. **Minimizing multivariate loss**: To optimize a specific multivariate performance measure, inspired by Joachims (2005), we propose to minimize a loss function $\Delta(\overline{y}^*, \overline{y})$ corresponding to this multivariate performance measure,

$$\min_{\mathbf{w}_j|_{j=1}^m} \Delta(\overline{y}^*, \overline{y}). \tag{8}$$

   Due to the complexity of the loss function, $\Delta(\overline{y}^*, \overline{y})$, instead of optimizing $\Delta(\overline{y}^*, \overline{y})$ directly, we seek its upper bound and optimize the upper bound. According to (2), the upper bound is obtained as follows,

$$\begin{aligned} & \sum_{j=1}^m \mathbf{w}_j^\top \Psi(\overline{x}^j, \overline{y}^*) \geq \sum_{j=1}^m \mathbf{w}_j^\top \Psi(\overline{x}^j, \overline{y}) \Rightarrow \Delta(\overline{y}^*, \overline{y}) \leq \\ & \sum_{j=1}^m \mathbf{w}_j^\top \left(\Psi(\overline{x}^j, \overline{y}^*) - \Psi(\overline{x}^j, \overline{y}) \right) + \Delta(\overline{y}^*, \overline{y}) \leq \\ & \max_{\overline{y}' \in \mathcal{Y}/\overline{y}} \left[\sum_{j=1}^m \mathbf{w}_j^\top \left(\Psi(\overline{x}^j, \overline{y}') - \Psi(\overline{x}^j, \overline{y}) \right) + \Delta(\overline{y}', \overline{y}) \right]. \end{aligned} \tag{9}$$

   Substituting the definition of $\Psi(\overline{\mathbf{x}}^j, \overline{y})$ and (3) to (9), we

can rewrite it as

$$\Delta(\overline{y}^*, \overline{y}) \leq \max_{\overline{y}' \in \mathcal{Y}/\overline{y}} \left[\mathbf{w}^\top \left(\sum_{j=1}^{m} \sum_{i=1}^{n} (y'_i - y_i) \Theta_j^\top \mathbf{x}_i^j \right) + \Delta(\overline{y}', \overline{y}) = \Delta(\overline{y}', \overline{y}) - \mathbf{w}^\top \boldsymbol{\pi}_{\overline{y}'} \right], \tag{10}$$

where

$$\boldsymbol{\pi}_{\overline{y}'} = \sum_{j=1}^{m} \sum_{i=1}^{n} (y_i - y'_i) \Theta_j^\top \mathbf{x}_i^j \in \mathbb{R}^d, \tag{11}$$

and the right hand side of (10) is an upper bound of $\Delta(\overline{y}^*, \overline{y})$. We further introduce a slack variable $\xi \geq 0$ to present the maximum value of the right hand side of (10), and minimize it to obtain an optimal performance measure,

$$\begin{aligned} &\min_{\mathbf{w}, \xi} \ \xi, \\ &s.t. \ \forall \overline{y}' \in \mathcal{Y}/\overline{y} : \ \xi \geq \Delta(\overline{y}', \overline{y}) - \mathbf{w}^\top \boldsymbol{\pi}_{\overline{y}'}, \ \xi \geq 0. \end{aligned} \tag{12}$$

The overall optimization function is obtained by combining the problems in (4), (6), and (12),

$$\begin{aligned} &\min_{\mathbf{w}, \xi} \left\{ \frac{1}{2} \mathbf{w}^\top \mathbf{w} + \frac{C_1}{2} \mathbf{w}^\top \Lambda \mathbf{w} + C_2 \xi \right\}, \\ &s.t. \ \forall \overline{y}' \in \mathcal{Y}/\overline{y} : \ \xi \geq \Delta(\overline{y}', \overline{y}) - \mathbf{w}^\top \boldsymbol{\pi}_{\overline{y}'}, \ \xi \geq 0, \end{aligned} \tag{13}$$

where C_1 and C_2 are the tradeoff parameters. In the objective of this problem, the first term is to reduce the complexity of the parameter vector of each view, the second term is to encourage the consistency of the responses of different views, and the third term is a slack variable to represent the upper boundary of the loss of the multivariate performance measure.

Problem optimization

To solve this problem, we employ the cutting-plane algorithm. Instead of using all the constraints in $\mathcal{Y}/\overline{y}$ to construct the optimization problem in (13), we only use an active set of constraints, $\mathcal{W}$, which contains a limited number of constraints in $\mathcal{Y}/\overline{y}$. In this algorithm, $\mathcal{W}$ and $\mathbf{w}$ are updated alternately.

Updating w When we have a given active set of constrain $\mathcal{W} \subseteq \mathcal{Y}/\overline{y}$, we replace $\mathcal{Y}/\overline{y}$ by $\mathcal{W}$ in (13), and obtain the following problem with regard to $\mathbf{w}$ and ξ,

$$\begin{aligned} &\min_{\mathbf{w}, \xi} \left\{ \frac{1}{2} \mathbf{w}^\top \mathbf{w} + \frac{C_1}{2} \mathbf{w}^\top \Lambda \mathbf{w} + C_2 \xi \right\}, \\ &s.t. \ \forall \overline{y}' \in \mathcal{W} : \ \xi \geq \Delta(\overline{y}', \overline{y}) - \mathbf{w}^\top \boldsymbol{\pi}_{\overline{y}'}, \ \xi \geq 0. \end{aligned} \tag{14}$$

The dual form of (14) is

$$\begin{aligned} \max_{\alpha_{\overline{y}'} | \overline{y}' : \overline{y}' \in \mathcal{W}} &\left\{ -\frac{1}{2} \sum_{\overline{y}', \overline{y}'' : \overline{y}', \overline{y}'' \in \mathcal{W}} \alpha_{\overline{y}'} \alpha_{\overline{y}''} \left(\boldsymbol{\pi}_{\overline{y}'}^\top (I + C_1 \Lambda)^{-1} \boldsymbol{\pi}_{\overline{y}''} \right) + \sum_{\overline{y}' : \overline{y}' \in \mathcal{W}} \alpha_{\overline{y}'} \Delta(\overline{y}', \overline{y}) \right\}, \\ &s.t. \ \forall \overline{y}' \in \mathcal{W} : \alpha_{\overline{y}'} \geq 0, \ C_2 \geq \sum_{\overline{y}' : \overline{y}' \in \mathcal{W}} \alpha_{\overline{y}'} \end{aligned} \tag{15}$$

where $\alpha_{\overline{y}'} | \overline{y}' : \overline{y}' \in \mathcal{W}$ are the Lagrange multipliers of the constrains in $\mathcal{W}$. This problem can be solved as a linear constrained quadratic programming problem. After the optimal $\alpha_{\overline{y}'} | \overline{y}' : \overline{y}' \in \mathcal{W}$ is solved, the optimal $\mathbf{w}$ can be recovered by $\mathbf{w} = (I + C_1 \Lambda)^{-1} \left(\sum_{\overline{y}' : \overline{y}' \in \mathcal{W}} \alpha_{\overline{y}'} \boldsymbol{\pi}_{\overline{y}'} \right)$.

Updating $\mathcal{W}$ With an updated $\mathbf{w}$, to approximate the upper bound of $\Delta(\overline{y}^*, \overline{y})$ in (10), we seek a $\overline{y}'$ to maximize the objective function of (10),

$$\overline{y}' = \arg \max_{\overline{y}'' \in \mathcal{Y}} \left\{ \Delta(\overline{y}'', \overline{y}) - \mathbf{w}^\top \boldsymbol{\pi}_{\overline{y}''} \right\}, \tag{16}$$

which is defined as the most violated constraint, and add it to the active label tuple set, $\mathcal{W} \leftarrow \mathcal{W} \cup \{\overline{y}'\}$. We use the method introduced in (Joachims 2005) to find the most violated constraint. The time complexity of this process for both F1-score and PRBEP is $O(n^2)$, and that of AUROC is $O(n \log n)$.

Iterative algorithm The proposed iterative algorithm is given in Algorithm 1. The convergence condition of this algorithm is reached when the most violated $\overline{y}'$ makes $\Delta(\overline{y}', \overline{y}) - \mathbf{w}^\top \boldsymbol{\pi}_{\overline{y}'} \leq \xi + \epsilon$, where ξ is the most current upper bound of the loss function, and ϵ is a convergence threshold. In most cases of our experiments, the convergence is reached within a maximum number of iterations (100 iterations in our experiments). For example, over a training set of the Handwritten digit data set, in 100 times of running of the algorithm with randomly initialized $\mathbf{w}$, a convergence rate of 92% is reported, indicating its convergence power. To scale this algorithm to large data sets, we can parallelize the steps relevant to the size of the data set, including the processes of calculating Λ, $\boldsymbol{\pi}_{\overline{y}'}$, and finding the constraint to maximize the objective of (16).

When updating $\mathbf{w}$, we solve a quadratic programming problem with the same number of variables as the standard SVM^{perf}, which is the number of current constraints. Compared to SVM^{perf}, the only complexity increase is from the increment of dimensionality when multi-view features are concatenated. However, if users concatenate the features from multi-views as one single-view and input it to SVM^{perf}, they will have the same complexity for its cutting-plane algorithm as our algorithm. Therefore, our algorithm shares the same complexity as SVM^{perf}.

Experiments

Here, we evaluate the proposed algorithm on four benchmark multi-view data sets for the problem of multivariate

Algorithm 1 Iterative multi-view learning algorithm for multivariate performance measure optimization (MVPO).

Input: Multi-view training data tuple $\overline{\mathbf{x}}^j|_{j=1}^m$ and class label tuple $\overline{y}$;
Input: The desired multivariate loss function $\Delta(\overline{y}', \overline{y})$.
Input: Tradeoff parameters C_1, C_2.
Initialize $\mathcal{W} = \emptyset$, $\mathbf{w}$, and calculate Λ as in (7);
repeat
  Find the most violated constraint $\overline{y}'$ by fixing $\mathbf{w}$ according to (16), and add $\overline{y}'$ to the active set, $\mathcal{W} \leftarrow \mathcal{W} \cup \{\overline{y}'\}$;
  Update $\mathbf{w}$ by fixing $\mathcal{W}$ and solving the problem in (14);
until A maximum iteration number is reached or convergency.
Output: $\mathbf{w}$.

performance measure optimization.

Data sets

The four benchmark data sets used include the handwritten digit data set (Van Breukelen et al. 1998), the CiteSeer scientific publication data set (Sen et al. 2008), the PASCAL VOC'07 image data set (Everingham et al. 2007), and the WebKB web page data set (Craven et al. 1998). The statistics of these data sets are given in Table 1. The six views of the digit images of the handwritten digit data set are six types of visual features, including Fourier coefficients of the character shapes, profile correlations, Karhunen-Love coefficients, pixel averages in 2×3 windows, Zernike moments, and morphological features. The three views of the publications in the CiteSeer data set are the textual view, and the two link views of inbound and outbound references. The two views of the images of the PASCAL VOC'07 data set are the visual view presented by the bag-of-words histogram of SIFT local features, and the textual features of user tags. The two views of the webpages of the WebKB data set are the textual view and the link view. These four data sets are useful to demonstrate the performance of the proposed problem setting, because the data of these data sets are both imbalanced and of multiple views. Therefore, the ideal classifiers on these data sets should be learned from multi-view data and evaluated by multivariate performance measures.

Table 1: The statistics of the four data sets.

Data set	#data points	#classes	#views
Handwritten digit	2,000	10	6
CiteSeer	3,312	6	3
PASCAL VOC'07	9,963	20	2
WebKB	1,051	2	2

Experimental protocol

To conduct the experiment, we equally split a data set to two subsets randomly, and used them as training and test sets respectively. Given a desired multivariate performance measure, we performed the proposed algorithm MVOP over the training data tuple and learned a classifier to optimize its corresponding multivariate loss function. Then we used the trained classifier to predict the class label tuple of the test data, and evaluated the prediction results using the desired multivariate performance measure. The desired multivariate performance measures are F1-score, PRBEP, and AUROC. The random splitting are repeated 10 times, and the mean macro-averages of the corresponding multivariate performance measure over all classes on the test set are reported as the results.

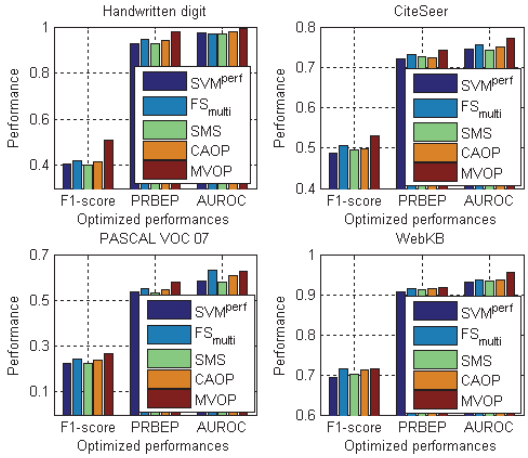

Figure 1: The mean macro-average multivariate performance comparison of multivariate performance optimization methods.

Comparison with multivariate performance optimization methods

We compared the proposed MVOP algorithm against four state-of-the-art single-view based multivariate performance optimization methods, including a SVM method for multivariate performance measures, SVM^{perf} (Joachims 2005), a feature selection method for multivariate performance measures, FS$_{multi}$ (Mao and Tsang 2013), a classifier adaptation method for multivariate performance measures, CAOP (Li, Tsang, and Zhou 2013), and a smoothing method for multivariate scores, SMS (Zhang, Saha, and Vishwanathan 2012). The choice of the view among multiple views and other parameters for baseline single-view methods is determined by cross-validation over the training sets, to optimize the performance of each method. The mean macro-average multivariate performance over the test sets of different methods is given in Figure 1. From this figure, we can see that in most cases, the proposed multi-view based performance optimization algorithm outperforms the single-view based methods. The only two exceptions are the AUROC performance over the PASCAL VOC'07 data set, and the F1-score performance over the WebKB data set, where FS$_{multi}$ outperforms MVOP slightly. The most significant improvements are made on the F1-score of the Handwritten digit data set, where only the average F1-score optimized by MVOP is higher than 0.5, while all other single-view based methods fail to exceed 0.5. This is a strong evidence for the

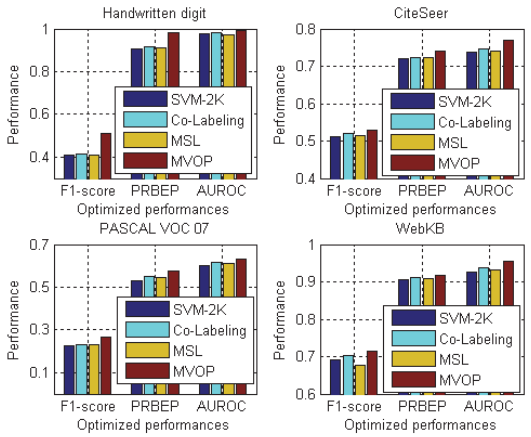

Figure 2: The mean macro-average multivariate performance comparisons of multi-view learning methods.

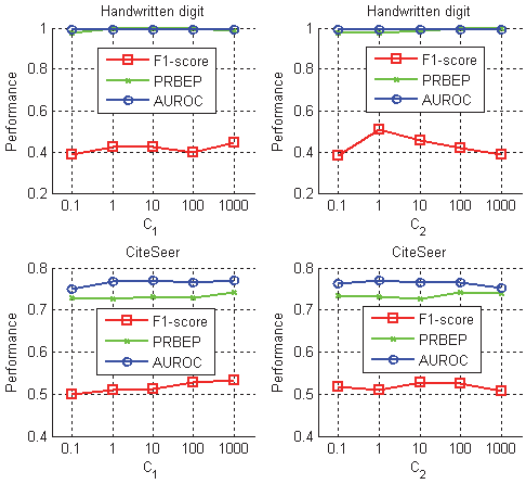

Figure 3: Performance curves of different values of tradeoff parameters on the Handwritten digit and the CiteSeer data sets.

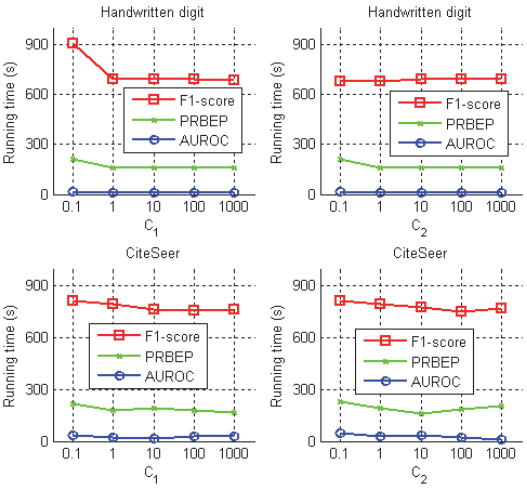

Figure 4: Training time (in seconds) of the proposed algorithm on the Handwritten digit data set and the CiteSeer data set.

effectiveness of the proposed algorithm, and it also reveals the necessity of exploring multiple views other than a single view to optimize the multivariate performance measures.

Comparison with multi-view learning methods

We also compared the proposed MVOP algorithm with three state-of-the-art multi-view learning algorithms, including a two view learning algorithm of SVM, SVM-2K (Farquhar et al. 2005), a multi-kernel learning based co-labeling algorithm for multi-view learning, Co-Labeling (Li et al. 2012), and a multi-view subspace learning algorithm, MSL (White et al. 2012). MSL is an unsupervised data representation algorithm. To run MSL, we first mapped the multi-view data to a shared sub space, and trained an SVM in the shared space for the purpose of classification. The mean macro-average multivariate performance of different multi-view learning methods is reported in Figure 2. Not surprisingly, all the compared traditional multi-view learning algorithms, which optimize the hinge or the squared hinge loss, cannot compete with the proposed MVOP algorithm which optimizes the target performance directly. This leads to our conclusion that using multi-view data does not lead to an optimal specific multivariate performance measure naturally, and learning performance-specific multi-view classifiers is important for the optimization of multivariate performance measures.

Sensitivity analysis of parameters

We analyzed the sensitivity of the proposed algorithm to the two tradeoff parameters, C_1 and C_2. The three optimized performance measures were plotted against different values of C_1 and C_2 on the Handwritten digit data set and the CiteSeer data set in Figure 3. The proposed algorithm, MVOP, is robust to the changes of both C_1 and C_2 values. In particular, the performance measures of PRBEP and AUROC are especially stable to C_1 and C_2 on the Handwritten digit data set, whereas the F1-score seems more sensitive to these parameters.

Training time analysis

We plot the training time of the proposed algorithm with different values of parameters C_1 and C_2 in Figure 4. As we can see from the figure, F1-score is the most time-consuming performance to optimize. It usually takes more than 600 seconds to perform the proposed algorithm to optimize the F1-score. Meanwhile, AUROC is the least time-consuming performance to optimize. This difference of training time is due to the difference in the process of finding the most validated constraint in each iteration. Moreover, we also observed that the training time is stable to the changes of the parameters, and an exception is the training time of optimizing F1-score with a small value of C_1, where about 900 seconds are consumed.

Conclusion

In this paper, we proposed the problem of multivariate performance optimization from multi-view data, and proposed a novel algorithm to solve it. Our algorithm learns and combines linear discriminant functions for different views to construct an overall multivariate mapping function for multi-view data. The learning problem is modeled as a minimization problem, where the function complexity is reduced, the view consistency is encouraged, and the upper bound of the multivariate loss is minimized. This problem is optimized by a cutting-plane algorithm. Experiments on four benchmark data sets demonstrate its advantages over the traditional multivariate performance measure optimization methods and the multi-view learning methods.

Acknowledgments

Research reported in this publication was supported by the King Abdullah University of Science and Technology (KAUST), the ARC Future Fellowship (Grant No. FT130100746), and the ARC grant (Grant No. LP150100671).

References

Craven, M.; McCallum, A.; PiPasquo, D.; Mitchell, T.; and Freitag, D. 1998. Learning to extract symbolic knowledge from the world wide web. Technical report, DTIC Document.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2007. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

Farquhar, J.; Hardoon, D.; Meng, H.; Shawe-taylor, J. S.; and Szedmak, S. 2005. Two view learning: Svm-2k, theory and practice. In *NIPS*, 355–362.

Joachims, T. 2005. A support vector method for multivariate performance measures. In *ICML*, 377–384. ACM.

Kelley, J. 1960. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics* 703–712.

Koyejo, O. O.; Natarajan, N.; Ravikumar, P. K.; and Dhillon, I. S. 2014. Consistent binary classification with generalized performance metrics. In *NIPS*, 2744–2752.

Li, W.; Duan, L.; Tsang, I. W.-H.; and Xu, D. 2012. Co-labeling: A new multi-view learning approach for ambiguous problems. In *ICDM*, 419–428.

Li, N.; Tsang, I. W.; and Zhou, Z.-H. 2013. Efficient optimization of performance measures by classifier adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(6):1370–1382.

Mao, Q., and Tsang, I. W.-H. 2013. A feature selection method for multivariate performance measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(9):2051–2063.

Mao, Q., Tsang, I. W., Gao, S., and Wang, L. 2015. Generalized Multiple Kernel Learning With Data-Dependent Priors. *IEEE Transactions on Neural Networks and Learning Systems* 26(6): 1134–1148.

Narasimhan, H.; Kar, P.; and Jain, P. 2015. Optimizing non-decomposable performance measures: A tale of two classes. In *ICML*.

Parambath, S. P.; Usunier, N.; and Grandvalet, Y. 2014. Optimizing f-measures by cost-sensitive classification. In *NIPS*, 2123–2131.

Sen, P.; Namata, G.; Bilgic, M.; Getoor, L.; Galligher, B.; and Eliassi-Rad, T. 2008. Collective classification in network data. *AI magazine* 29(3):93.

Sun, S. 2013. A survey of multi-view machine learning. *Neural Computing and Applications* 23(7-8):2031–2038.

Van Breukelen, M.; Duin, R. P.; Tax, D. M.; and Den Hartog, J. 1998. Handwritten digit recognition by combined classifiers. *Kybernetika* 34(4):381–386.

Wang, J. J.-Y., and Gao, X. 2015. Partially labeled data tuple can optimize multivariate performance measures. In *CIKM*, 1915–1918. ACM.

White, M.; Zhang, X.; Schuurmans, D.; and Yu, Y.-l. 2012. Convex multi-view subspace learning. In *NIPS*, 1673–1681.

Xu, X.; Li, W.; Xu, D.; and Tsang, I. 2015. Co-labeling for multi-view weakly labeled learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP(99):1–1.

Xu, C.; Tao, D.; and Xu, C. 2013. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.

Xu, C.; Tao, D.; and Xu, C. 2015. Multi-view intact space learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP(99):1–1.

Yang, J.; Ding, Z.; Guo, F.; Wang, H.; and Hughes, N. 2015. A novel multivariate performance optimization method based on sparse coding and hyper-predictor learning. *Neural Networks* 71:45–54.

Zhang, X.; Saha, A.; and Vishwanathan, S. V. N. 2012. Smoothing multivariate performance measures. *J. Mach. Learn. Res.* 13(1):3623–3680.