Optimizing Data Transformation for Binary Classification

Kangrok Oh, Kar-Ann Toh, and Zhengguo Li

***Abstract*—In this paper, we propose to optimize a data transformation matrix and study its impact on binary classification. Based on the area above the receiver operating characteristics curve (AAC) minimization with data transformation, we optimize alternatingly between the data transformation matrix and the weighting parameter vector. Some experimental results on 16 binary data sets acquired from the UCI machine learning repository are observed and discussed. Classification accuracy and ranking value averaged from 10 runs of stratified 10-fold cross-validation are adopted as performance indicators. The proposed method shows encouraging results based on these two performance indicators. In addition, it is shown that most of the performance comparisons are statistically significant.**

***Index Terms*—Data transformation, machine learning, pattern classification, receiver operating characteristics curve.**

I. Introduction

The receiver operating characteristics (ROC) curve is a plot of true positive rates over false positive rates with respect to various operating threshold values. Since the ROC curve provides an overall performance of a pattern classifier, it has been widely utilized as a performance measure for pattern classification tasks [1]-[5]. In a qualitative manner, a classifier is considered to perform well when its ROC curve is drawn close to the upper-left corner. Apart from this qualitative measure, the area under the ROC curve (AUC) provides a quantitative measure regarding the ROC performance [6].

According to [6], the AUC can be computed based on the sum of entire pairwise comparisons between data features of opposite categories. The value of the AUC is identical to that of the Wilcoxon-Mann-Whitney statistic [7] which uses a zero-one step loss function for the comparison. In [8], it has been shown that a quadratic approximation to the zero-one step loss function can be effectively adopted for analytic AUC maximization. Furthermore, it has been shown in [9] that several existing classifiers such as least squares estimation (LSE) [10], Fishers' linear discriminants [11], total error rate (TER) minimization [12] can be linked together under a transformed AUC (called TAUC) framework based on the quadratic approximation and a data transformation.

Based on this interesting link among the classifiers which hinges on a data transformation matrix, our motivations for this work can be enumerated as follows:

1) The TAUC formulation [9] utilizes only diagonal components of the data transformation (scaling) matrix. An investigation into data transformation using a full matrix would verify whether such additional transformation can help classification generalization.
2) The novel classifier [9] utilizes an ad-hoc and random settings to go beyond existing classifier platform. Relevant classifier setting adopting an optimization process is apparently more appealing than the ad-hoc and random setup.

Particularly, we aim to optimize the data transformation under AUC criterion and study its impact on binary classification. The contribution of this work can be summarized as follows:

1) We provide an optimal data transformation matrix under AUC maximization (AAC minimization) criterion.
2) We show the impact of such optimal solution on binary classification via experimentation on 16 data sets from the UCI machine learning repository.

The paper is organized as follows. In Section II, background information on a linear parametric model and the AAC minimization criterion is presented. An alternating AAC minimization methodology is proposed in Section III. Experimental results and analysis are provided in Section IV. Finally, some concluding remarks are given in Section V.

II. Preliminaries

A. Linear Parametric Model

Given a feature vector $\mathbf{x} \in \mathbb{R}^{D\times 1}$, a linear parametric model adopting nonlinear feature expansion can be written as

$$g(\boldsymbol{\alpha}, \mathbf{x}) = \sum_{j=1}^{K} \alpha_j p_j(\mathbf{x}) = \mathbf{p}(\mathbf{x})^T \boldsymbol{\alpha}, \tag{1}$$

where $\mathbf{p}(\mathbf{x}) \in \mathbb{R}^{K\times 1}$ denotes a nonlinear feature expansion vector (such as polynomial expansion) and $\boldsymbol{\alpha} \in \mathbb{R}^{K\times 1}$ denotes a parameter vector consisting of feature weighting coefficients.

B. AAC Minimization Criterion

An equivalent way to maximize the AUC is to minimize the AAC which is the area above the ROC curve. In other words, AAC = 1-AUC since they are normalized quantities. The quadratic approximated AAC criterion function with weight decay regularization from [8] can be expressed as

$$J(\boldsymbol{\alpha}, \mathbf{p}_{j,i}) = \frac{b}{2}\boldsymbol{\alpha}^T\boldsymbol{\alpha} + \frac{1}{2M^+M^-}\sum_{i=1}^{M^+}\sum_{j=1}^{M^-}\left[\mathbf{p}_{j,i}^T\boldsymbol{\alpha} + \eta\right]^2, \tag{2}$$

where $\mathbf{p}_{j,i} = \mathbf{p}(\mathbf{x}_j^-) - \mathbf{p}(\mathbf{x}_i^+) \in \mathbb{R}^{K\times 1}$ is a difference vector

Manuscript received June 13, 2015; revised November 15, 2015. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (Grant number: NRF-2012R1A1A2042428).

Kangrok Oh and Kar-Ann Toh are with School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea (e-mail: kangrokoh@ yonsei.ac.kr, katoh@ yonsei.ac.kr).

Zhengguo Li is with the Institute for Infocomm Research, Singapore 119613, Singapore (e-mail: ezgli@i2r.a-star.edu).

DOI: 10.7763/IJCTE.2017.V9.1102

between each sample in positive class and each sample in negative class, $\boldsymbol{\alpha} \in \mathbb{R}^{K\times 1}$ is a weighting coefficient vector to be estimated, η is an offset value, and M^+ and M^- are the number of samples in positive and negative samples, respectively.

III. Proposed Method

A. Problem Formulation

Consider a data transformation defined as

$$\mathbf{p}(\mathbf{x}) = \mathbf{A}\mathbf{p}(\mathbf{x}), \tag{3}$$

where $\mathbf{p}(\mathbf{x}) \in \mathbb{R}^{K\times 1}$ is a nonlinear feature expansion vector (such as polynomial expansion), $\mathbf{A} \in \mathbb{R}^{K\times K}$ is a data transformation matrix to be optimized, $\mathbf{p}(\mathbf{x}) \in \mathbb{R}^{K\times 1}$ is a transformed feature vector, and K is the feature dimension.

This data transformation can be applied to the quadratic approximated AAC criterion function with weight decay regularization [8] as

$$J\left(\boldsymbol{\alpha}, \mathbf{A}, \mathbf{p}_{j,i}\right) = \frac{b}{2}\boldsymbol{\alpha}^T\boldsymbol{\alpha} + \frac{1}{2M^+M^-}\sum_{i=1}^{M^+}\sum_{j=1}^{M^-}\left[\mathbf{p}_{j,i}^T\mathbf{A}^T\boldsymbol{\alpha} + \eta\right]^2. \tag{4}$$

In order to rewrite the double summations in (4) into a single summation form for ease of algebraic manipulation, a new index k is defined such that $k = M^-(i-1) + j$ for $i = 1,\ldots,M^+$ and $j = 1,\ldots,M^-$. Then, each pair of i, j index corresponds to a single value of k which falls in the range $\{1,2,\ldots,N\}$ where $N = M^+ \times M^-$. By representing $\mathbf{p}_{j,i}$ as $\mathbf{q}_k$, (4) can be re-written as

$$J\left(\boldsymbol{\alpha}, \mathbf{A}, \mathbf{q}_k\right) = \frac{b}{2}\boldsymbol{\alpha}^T\boldsymbol{\alpha} + \frac{1}{2N}\sum_{k=1}^{N}\left[\mathbf{q}_k^T\mathbf{A}^T\boldsymbol{\alpha} + \eta\right]^2. \tag{5}$$

Next, by stacking the $\mathbf{q}_k$ vectors into a matrix $\mathbf{Q}$, a matrix form of (5) can be re-written as

$$J\left(\boldsymbol{\alpha}, \mathbf{A}, \mathbf{Q}\right) = \frac{b}{2}\boldsymbol{\alpha}^T\boldsymbol{\alpha} + \frac{1}{2N}\boldsymbol{\alpha}^T\mathbf{A}\mathbf{Q}\mathbf{Q}^T\mathbf{A}^T\boldsymbol{\alpha} + \frac{\eta}{N}\mathbf{1}^T\mathbf{Q}^T\mathbf{A}^T\boldsymbol{\alpha} + \frac{\eta^2}{2}. \tag{6}$$

where $\mathbf{Q} = \left[\mathbf{q}_1,\ldots,\mathbf{q}_N\right] \in \mathbb{R}^{K\times N}$, and $\mathbf{1} = \left[1,\ldots,1\right]^T \in \mathbb{R}^{N\times 1}$ is a vector consisting of only '1' values. We will use (6) as the objective function.

B. Proposed Alternating AAC Minimization

In this section, we propose to minimize the AAC criterion function with weight decay regularization and data transformation (6), with respect to $\boldsymbol{\alpha}$ and $\mathbf{A}$ alternatingly.

The optimality condition for $\mathbf{A}$ which minimizes (6) is to solve for $\mathbf{A}$ when

$$\frac{\partial J\left(\boldsymbol{\alpha}, \mathbf{A}, \mathbf{Q}\right)}{\partial \mathbf{A}} = 0, \tag{7}$$

Which is equivalent to solving

$$\frac{1}{N}\boldsymbol{\alpha}\boldsymbol{\alpha}^T\mathbf{A}\mathbf{Q}\mathbf{Q}^T + \frac{\eta}{N}\boldsymbol{\alpha}\mathbf{1}^T\mathbf{Q}^T = 0. \tag{8}$$

To solve (8) with respect to $\mathbf{A}$, two matrix inverse operations are required. Here, we utilize a small regularization constant (such as 10^{-4}) in order to prevent singularity condition for the inverse operations as

$$\mathbf{A} = -\eta\left(b\mathbf{I} + \boldsymbol{\alpha}\boldsymbol{\alpha}^T\right)^{-1}\boldsymbol{\alpha}\mathbf{1}^T\mathbf{Q}^T\left(b\mathbf{I} + \mathbf{Q}\mathbf{Q}^T\right)^{-1}, \tag{9}$$

where $\mathbf{I} \in \mathbb{R}^{K\times K}$ denotes an identity matrix.

In (9), the matrix inverse operation $\left(b\mathbf{I} + \boldsymbol{\alpha}\boldsymbol{\alpha}^T\right)^{-1}$ includes a singular matrix. In order to calculate the inverse operation under smaller dimension among the two, we adopt a matrix identity $\left(\mathbf{A} + \mathbf{B}\mathbf{B}^T\right)^{-1}\mathbf{B} = \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{I} + \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\right)^{-1}$ [13] to $\left(b\mathbf{I} + \boldsymbol{\alpha}\boldsymbol{\alpha}^T\right)^{-1}\boldsymbol{\alpha}$. Then, (9) can be re-written as

$$\begin{aligned}\mathbf{A} &= -\eta\left(b\mathbf{I}\right)^{-1}\boldsymbol{\alpha}\left(\mathbf{I} + \boldsymbol{\alpha}^T\left(b\mathbf{I}\right)^{-1}\boldsymbol{\alpha}\right)^{-1} \times \mathbf{1}^T\mathbf{Q}^T\left(b\mathbf{I} + \mathbf{Q}\mathbf{Q}^T\right)^{-1} \\ &= -\eta\boldsymbol{\alpha}\left(b + \boldsymbol{\alpha}^T\boldsymbol{\alpha}\right)^{-1}\mathbf{1}^T\mathbf{Q}^T\left(b\mathbf{I} + \mathbf{Q}\mathbf{Q}^T\right)^{-1}.\end{aligned} \tag{10}$$

The optimality condition for $\boldsymbol{\alpha}$ which minimizes (6) is

$$\frac{\partial J\left(\boldsymbol{\alpha}, \mathbf{A}, \mathbf{Q}\right)}{\partial \boldsymbol{\alpha}} = 0, \tag{11}$$

and this implies

$$b\boldsymbol{\alpha} + \frac{1}{N}\mathbf{A}\mathbf{Q}\mathbf{Q}^T\mathbf{A}^T\boldsymbol{\alpha} + \frac{\eta}{N}\mathbf{A}\mathbf{Q}\mathbf{1} = 0. \tag{12}$$

The solution for $\boldsymbol{\alpha}$ in (12) is

$$\boldsymbol{\alpha} = \frac{-\eta}{N}\left(b\mathbf{I} + \frac{1}{N}\mathbf{A}\mathbf{Q}\mathbf{Q}^T\mathbf{A}^T\right)^{-1}\left(\mathbf{A}\mathbf{Q}\mathbf{1}\right). \tag{13}$$

The proposed minimization optimizes both data transform matrix $\mathbf{A}$ and coefficient vector $\boldsymbol{\alpha}$. This is different from the AAC minimization in [8], and the TAAC minimization in [9]. The proposed alternating minimization algorithm is summarized as a pseudocode in Algorithm 1. When t is equal to 1, the solution for $\boldsymbol{\alpha}$ to (11) is exactly same as the AAC solution.

Algorithm 1 Pseudocode for the proposed alternating AAC minimization method

```
Input: A matrix $\mathbf{Q}$ consisting of pre-defined feature vectors
       Number of iterations $T$
Output: A weighting coefficient vector $\boldsymbol{\alpha}$
        A data transformation matrix $\mathbf{A}$

for $t = 1$ to $T$ do
    if $t = 1$ then
        $\mathbf{A} \Leftarrow \mathbf{I}$                {Initialization}
    else
        $\mathbf{A} \Leftarrow -\eta\boldsymbol{\alpha}\left(b + \boldsymbol{\alpha}^T\boldsymbol{\alpha}\right)^{-1}\mathbf{1}^T\mathbf{Q}^T\left(b\mathbf{I} + \mathbf{Q}\mathbf{Q}^T\right)^{-1}$    {Optimal $\mathbf{A}$}
    end if
    $\boldsymbol{\alpha} \Leftarrow \frac{-\eta}{N}\left(b\mathbf{I} + \frac{1}{N}\mathbf{A}\mathbf{Q}\mathbf{Q}^T\mathbf{A}^T\right)^{-1}\left(\mathbf{A}\mathbf{Q}\mathbf{1}\right)$    {Optimal $\boldsymbol{\alpha}$}
end for
```

At testing stage, the class label of a nonlinearly expanded test feature vector $\mathbf{p}(\mathbf{x}_t) \in \mathbb{R}^{K \times 1}$ can be estimated using

$$cls(\mathbf{p}(\mathbf{x}_t)) = \begin{cases} 0 & , \text{ if } g(\boldsymbol{\alpha}, \mathbf{p}(\mathbf{x}_t)) < \tau \\ 1 & , \text{ if } g(\boldsymbol{\alpha}, \mathbf{p}(\mathbf{x}_t)) \ge \tau \end{cases}. \qquad (14)$$

where τ is an optimal threshold based on TER minimization [9], and $g(\boldsymbol{\alpha}, \mathbf{p}(\mathbf{x}_t)) = \mathbf{p}(\mathbf{x}_t)^T \boldsymbol{\alpha}$.

IV. Experiments

A. Database and Preprocessing

The databases utilized in our experiments consist of 16 binary data sets obtained from the UCI machine learning repository [14]. The data has categorical, integer, and real attributes. The categorical attributes are changed to values between 0 and 1 with equal spacing except for Monk-1 database. For Monk-1 database, the first two attributes are represented within the range $[0,10]$ according to [10]. The real and integer values are scaled to the range within $[0,1]$. For samples with don't care attributes in shuttle database, every possible values are generated. Samples with missing values are excluded in WBC, credit, mushroom, and WPBC databases. Table I shows a summary of the utilized databases.

B. Experimental Setups

In order to assess the classification performance of the proposed AAAC minimization, several well-known classifiers such as LSE [10], TER minimization [12], Fisher's linear discriminant (FLD) [11], AAC minimization [8], and novel classifier [9] are adopted for comparison. For TER minimization, two different settings, namely TERa and TERb, are applied according to [12]. A reduced multivariate polynomial (RM) model [10] is adopted as a nonlinear feature expansion function for LSE, TERa, TERb, AAC, novel classifier, and AAAC minimization. For order of RM model, we select from a set of integer values within the set $\{1, \ldots, 10\}$ based on cross-validation using only training data.

TABLE I: Summary of the 16 Binary Data Sets Obtained from UCI Macine Learning Repository

Database	Data Type	Data Dimension	Number of Classes	Number of Given Samples	Number of Utilized Samples	(m[20]/M[21])[22]	Note
Shuttle[1]	C[17]	6	2	15	278	0.9172	Samples with 'don't care' values are manipulated by considering every possible values
Liver[2]	C, I[18], R[19]	6	2	345	345	0.7250	-
Monk-1[3]	C	6	2	124	124	1.0000	The first two attributes are scaled within the range [0,10]
Monk-2[4]	C	6	2	169	169	0.6095	-
Monk-3[5]	C	6	2	122	122	0.9677	-
Pima[6]	I, R	8	2	768	768	0.5360	-
Tic-Tac-Toe[7]	C	9	2	958	958	0.5304	-
WBC[8]	I	9	2	699	683	0.5383	16 samples with missing values are excluded
Heart[9]	C, R	13	2	270	270	0.8000	-
Credit[10]	C, I, R	15	2	690	653	0.8291	37 samples with missing values are excluded
Voting[11]	C	16	2	435	435	0.6292	-
Mushroom[12]	C	22	2	8124	5644	0.6181	2480 samples with missing values are excluded
WDBC[13]	R	30	2	569	569	0.5938	-
WPBC[14]	R	30	2	198	194	0.3108	4 samples with missing values are excluded
Ionosphere[15]	I, R	34	2	351	351	0.5600	-
Sonar[16]	R	60	2	208	208	0.8739	-

[1] Shuttle: Shuttle Landing Control Data Set [2] Liver: Liver Disorders Data Set [3] Monk-1: MONK's Problems Data Set (Monk-1 Subset)
[4] Monk-2: MONK's Problems Data Set (Monk-2 Subset) [5] Monk-3: MONK's Problems Data Set (Monk-3 Subset)
[6] Pima: Pima Diabetes Database [7] Tic-Tac-Toe: Tic-Tac-Toe Endgame Data Set [8] WBC: Breast Cancer Wisconsin (Original) Data Set
[9] Heart: Statlog (Heart) Data Set [10] Credit: Credit Approval Data Set [11] Voting: Congressional Voting Records Data Set
[12] Mushroom: Mushroom Data Set [13] WDBC: Breast Cancer Wisconsin (Diagnostic) Data Set
[14] WPBC: Breast Cancer Wisconsin (Prognostic) Data Set [15] Ionosphere: Ionosphere Data Set
[16] Sonar: Connectionist Bench (Sonar, Mines vs. Rocks) Data Set [17] C: Categorical [18] I: Integer [19] R: Real
[20] m: Number of samples in a class with smaller number of samples [21] M: Number of samples in a class with larger number of samples
[22] m/M: the value is regarding class data distribution. A larger value of m/M corresponds to more balanced data distribution

For regularization, a constant value of 10^{-4} is adopted for all compared classifiers except for TERb. For TERb, a tuning choice within $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ is adopted according to [12]. We set 0.5 as the threshold value for LSE, TERa, and TERb, and an optimal threshold minimizing TER [9] is adopted for FLD, novel classifier, AAC minimization, and the proposed AAAC minimization as in [9]. The number of iterations (T) of the proposed AAAC minimization is set to 3. For the novel classifier, we utilized six different settings based on translation vector type and random value ranges. The parameters of the proposed AAAC minimization and other classifiers utilized for performance comparisons are summarized in Table II.

The classification accuracy given by $(tp+tn)/(M^+ + M^-)$ is adopted as the performance measure in the experiments where tp and tn denote true positive and true negative respectively. The accuracy performance is reported in terms of the average accuracy value obtained from 10 runs of stratified 10-fold cross validation tests. The hyper parameter values are selected based on a single run of stratified 10-fold cross-validation using only the training set. Additionally, we report results from statistical tests such as Friedman and Nemenyi tests [15] to see if the difference in performance is of statistical significance.

C. Results

Table III shows the average classification accuracy and ranking values based on 10 runs of stratified 10-fold cross-validation tests. Novel classifier shows the best average classification accuracy due to its adoption of the best among several solutions which is analogous to running of several existing classifiers. TERa shows the second best average classification accuracy attributed to its capability for class-specific normalization. TERb shows worse average classification accuracy than TERa due to over-training. The proposed AAAC minimization shows the third best average classification accuracy. Comparing with the original AAC minimization, AAAC minimization shows minor improvement in terms of average classification accuracy. The AAC based classifiers such as AAC and AAAC minimization and novel classifier show better average classification accuracy than LSE and FLD.

TABLE II: PARAMETER SETTINGS FOR THE CLASSIFIERS

Parameter	Meaning	Classifier						
		LSE	FLD	TERa	TERb	AAC	Novel	Proposed
r	RM order	$r \in \{1, 2, \ldots, 10\}$						
b	Regularization constant	10^{-4}	10^{-4}	10^{-4}	B[2]	10^{-4}	10^{-4}	10^{-4}
τ	Threshold	0.5	opt[1]	0.5	0.5	opt	opt	opt
η	Offset	-	-	0.5	0.5	0.5	0.5	0.5
$[l,u]$	l = the lower bound for random value u = the upper bound for random value	-	-	-	-	-	0 $\{1,10,1000\}$	-
$rand$	Translation parameter vector type	-	-	-	-	-	{same[3], diff[4]}	-
L	The number of repetitions for random setting	-	-	-	-	-	10	-
T	The number of iterations	-	-	-	-	-	-	3

[1] opt: optimal threshold [2] B: $B=\{10^{-1},10^{-2},10^{-3},10^{-4}\}$ [3] same: translation vector consisting of the same random value

[4] diff: translation vector consisting of different random value

TABLE III: CLASSIFICATION ACCURACY AND RANKING[1] AVERAGED FROM 10 RUNS OF STRATIFIED 10-FOLD CROSS-VALIDATION

Database	Classifier						
	LSE	FLD	TERa	TERb	AAC	Novel	AAAC
Shuttle	0.9605(3)	0.9563(4)	**0.9696(1.5)**	0.9540(7)	0.9556(5.5)	**0.9696(1.5)**	0.9556(5.5)
Liver	0.7150(4)	0.7172(2)	0.7150(4)	**0.7186(1)**	0.7145(6.5)	0.7150(4)	0.7145(6.5)
Monk-1	0.9467(6)	0.9774(5)	0.9874(4)	0.8751(7)	**0.9983(2)**	**0.9983(2)**	**0.9983(2)**
Monk-2	0.7603(3)	0.7491(6)	**0.7623(1.5)**	0.7302(7)	0.7522(5)	**0.7623(1.5)**	0.7523(4)
Monk-3	0.9096(5.5)	0.9096(7)	0.9096(5.5)	0.9118(4)	**0.9141(2)**	**0.9141(2)**	**0.9141(2)**
Pima	**0.7716(2)**[2]	0.7619(5)	**0.7716(2)**	0.7702(4)	0.7575(6.5)	**0.7716(2)**	0.7575(6.5)
Tic-Tac-Toe	0.9833(3.5)	0.9833(3.5)	0.9833(3.5)	**0.9833(1)**	0.9833(6.5)	0.9833(3.5)	0.9833(6.5)
WBC	0.9693(6)	0.9703(5)	**0.9732(1.5)**	0.9677(7)	0.9731(3.5)	**0.9732(1.5)**	0.9731(3.5)
Heart	0.8407(2)	0.8393(3)	0.8356(7)	**0.8415(1)**	0.8381(5)	0.8381(5)	0.8381(5)
Credit	0.8653(2)	0.8650(5)	0.8650(3.5)	**0.8677(1)**	0.8647(6.5)	0.8650(3.5)	0.8647(6.5)
Voting	0.9544(4)	0.9541(6)	**0.9544(1.5)**	0.9385(7)	0.9544(4)	**0.9544(1.5)**	0.9544(4)
Mushroom	**1.0000(2.5)**	**1.0000(2.5)**	**1.0000(2.5)**	0.9925(7)	0.9986(5.5)	**1.0000(2.5)**	0.9986(5.5)
WDBC	0.9554(7)	0.9620(5)	0.9708(4)	0.9589(6)	**0.9714(2)**	**0.9714(2)**	**0.9714(2)**
WPBC	**0.7926(2)**	0.7160(7)	**0.7926(2)**	0.7635(4.5)	0.7270(6)	**0.7926(2)**	0.7635(4.5)
Ionosphere	0.8654(5)	0.8666(4)	0.8716(2.5)	**0.8765(1)**	0.8631(6.5)	0.8716(2.5)	0.8631(6.5)
Sonar	0.7444(6)	0.7396(7)	0.7567(5)	**0.7893(1)**	0.7865(3)	0.7865(3)	0.7865(3)
Average	0.8772(3.97)	0.8730(4.81)	0.8824(3.22)	0.8712(4.16)	0.8783(4.75)	**0.8854(2.50)**	0.8806(4.59)

[1] Ranking: when ranking of multiple classifiers are the same, average ranking is provided.

[2] A bold letter denotes the best accuracy and the best ranking.

In terms of average ranking, the compared classifiers can be ranked in descending order as {novel classifier, TERa, LSE, TERb, AAAC, AAC, FLD}. Here we note that LSE shows the third best average ranking unlike the average classification accuracy comparison. This is because little difference in accuracy performance can result in relatively big difference in ranking (see results on Tic-Tac-Toe database in Table III as an extreme example).

TABLE IV: FRIEDMAN TEST RESULTS USING CLASSIFICATION ACCURACY

Assumption	p	$p < 0.01$
$p = 0.01$	7.4125×10^{-25}	Reject null

Table IV shows the results from a Friedman test using classification accuracy. Since the results reject null hypothesis that all algorithms are equivalent at p < 0.01, Nemenyi test can be further performed to observe the difference among the algorithms. Fig. 1 shows the results from Nemenyi test. As illustrated in Fig. 1, novel classifier and TERa show better classification accuracy performance than other classifiers.

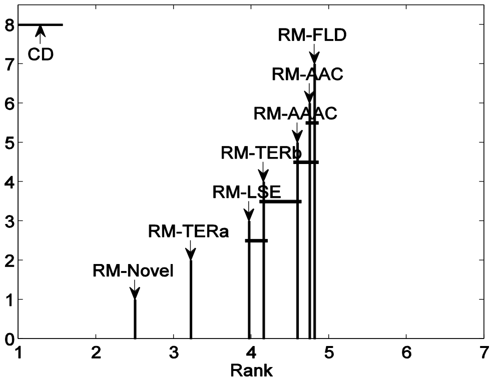

Fig. 1. Nemenyi test results using classification accuracy.

We observe that AAAC minimization is not suitable to use as a stand-alone classifier in view of its average classification accuracy and ranking. As for future work, we propose to combine it with novel and other existing classifiers to form a stronger classifier.

V. Conclusion

In this paper, we proposed to optimize a data transformation matrix for binary classification. An alternating framework based on the area above the ROC curve minimization with respect to a data transformation matrix and a weighting coefficient vector was presented. Our experimental results and analysis on 16 UCI machine learning repository databases showed encouraging results in terms of average classification accuracy. In addition, we showed that most of the performance comparisons are significantly different in statistical sense.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (Grant number: NRF-2012R1A1A2042428).

References

[1] A. K. Jain, S. Prabhakar, L. Hong, and S. Pankanti, "Filterbank-based fingerprint matching," *IEEE Trans. on Image Processing*, vol. 9, no. 5, pp. 846–859, May 2000.

[2] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *Proc. 1994 IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, Seattle, 1994, pp. 84–91.

[3] S. Gefen *et al*., "ROC analysis of ultrasound tissue characterization classifiers for breast cancer diagnosis," *IEEE Trans. on Medical Imaging*, vol. 22, no. 2, pp. 170–177, Feb. 2003.

[4] J. Shiraishi *et al*., "Computer-aided diagnosis to distinguish benign from malignant solitary pulmonary nodules on radiographs: ROC analysis of radiologists performance - initial experience," *Radiology*, vol. 227, no. 2, pp. 469–474, May 2003.

[5] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. 2010 IEEE Conf. Computer Vision and Pattern Recognition*, San Francisco, 2010, pp. 1975–1981.

[6] C. Cortes and M. Mohri, "AUC optimization vs. error rate minimization," *Advances in Neural Information Processing Systems*, vol. 16, no. 16, pp. 313–320, 2004.

[7] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, April 1982.

[8] K.-A. Toh, J. Kim, and S. Lee, "Maximizing area under ROC curve for biometric scores fusion," *Pattern Recognition*, vol. 41, no. 11, pp. 3373–3392, Nov. 2008.

[9] K.-A. Toh and G.-C. Tan, "Exploiting the relationships among several binary classifiers via data transformation," *Pattern Recognition*, vol. 47, no. 3, pp. 1509–1522, March 2014.

[10] K.-A. Toh, Q.-L. Tran, and D. Srinivasan, "Benchmarking a reduced multivariate polynomial pattern classifier," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 740–755, June 2004.

[11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. John Wiley & Sons, 2001.

[12] K.-A. Toh and H.-L. Eng, "Between classification-error approximation and weighted least-squares learning," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 658–669, April 2008.

[13] K. B. Petersen and M. S. Pedersen, *The Matrix Cookbook*, Nov. 2012.

[14] M. Lichman, *UCI Machine Learning Repository*, 2013.

[15] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, Dec. 2006.

Kangrok Oh received the B.S. and M.S. degrees in electrical and electronic engineering from Yonsei Uni-versity, Seoul, Korea, in 2010 and 2012 respectively. He is currently a Ph.D. candidate in the School of Electrical and Electronic Engineering at Yonsei University. His research interests include pattern recognition and biometrics.

Kar-Ann Toh is a professor in the School of Electrical and Electronic Engineering at Yonsei University, South Korea. He received the PhD degree from Nanyang Technological University (NTU), Singapore.

He was affiliated with Institute for Infocomm Research in Singapore from 2002 to 2005 prior to his current appointment in Korea. His research interests include biometrics, pattern classification, optimization and neural networks.

Dr. Toh is currently an associate editor of *IEEE Transactions on Information Forensic* and *Security, Pattern Recognition Letters* and *IET Biometrics*.

Zhengguo Li received the B.Sci. and M.Eng. from Northeastern University, Shenyang, China, in 1992 and 1995, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2001.

His current research interests include computational photography, mobile imaging, video processing & delivery, QoS, hybrid systems, and chaotic secure communication. He has co-authored one monograph, more than 160 journal/conference papers, and six granted patents, including normative technologies on scalable extension of H.264/AVC. He has been actively involved in the development of H.264/AVC and HEVC since 2002. He had three informative proposals adopted by the H.264/AVC and three normative proposals adopted by the HEVC. Currently, he is with the Agency for Science, Technology and Research, Singapore. He is an elected Technical Committee of the IEEE Visual Signal Processing and Communication. He served as a technical chair of IEEE ICIEA in 2010, a General Chair of IEEE ICIEA in 2011, a technical brief co-chair of SIGGRAPH Asia in 2012, a general co-chair of CCDC in 2013, and the workshop chair of IEEE ICME in 2013. He was an associate editor of *IEEE Signal Processing Letters* since 2014. He was an invited lecturer by the 2011 IEEE Signal Processing Summer School on 3-D and High Definition/High Contrast Video Processing Systems and a distinguished invited lecturer by IEEE INDIN in 2012.