Received March 10, 2016, accepted March 27, 2016, date of publication April 7, 2016, date of current version April 21, 2016.

Digital Object Identifier 10.1109/ACCESS.2016.2551727

Regularization and Kernelization of the Maximin Correlation Approach

TAEHOON LEE[1], TAESUP MOON[2], (Member, IEEE), SEUNG JEAN KIM[3], (Member, IEEE), AND SUNGROH YOON[1], (Senior Member, IEEE)

[1]Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, South Korea
[2]Department of Information and Communication Engineering, Daegu-Gyeongbuk Institute of Science and Technology, Daegu 42988, South Korea
[3]Citi Capital Advisors, New York, NY 10013, USA

Corresponding author: S. Yoon (sryoon@snu.ac.kr)

This work was supported in part by the National Research Foundation of Korea under Grant 2011-0009963, in part by the Institute for Information Communications Technology Promotion under Grant B0101-16-0307 (Basic Software Research in Human-level Lifelong Machine Learning), and in part by the Korea Health Industry Development Institute under Grant HI14C3405030014.

ABSTRACT Robust classification becomes challenging when each class consists of multiple subclasses. Examples include multi-font optical character recognition and automated protein function prediction. In correlation-based nearest-neighbor classification, the maximin correlation approach (MCA) provides the worst-case optimal solution by minimizing the maximum misclassification risk through an iterative procedure. Despite the optimality, the original MCA has drawbacks that have limited its wide applicability in practice. That is, the MCA tends to be sensitive to outliers, cannot effectively handle nonlinearities in datasets, and suffers from having high computational complexity. To address these limitations, we propose an improved solution, named regularized MCA (R-MCA). We first reformulate MCA as a quadratically constrained linear programming (QCLP) problem, incorporate regularization by introducing slack variables in the primal problem of the QCLP, and derive the corresponding Lagrangian dual. The dual formulation enables us to apply the kernel trick to R-MCA, so that it can better handle nonlinearities. Our experimental results demonstrate that the regularization and kernelization make the proposed R-MCA more robust and accurate for various classification tasks than the original MCA. Furthermore, when the data size or dimensionality grows, R-MCA runs substantially faster by solving either the primal or dual (whichever has a smaller variable dimension) of the QCLP. The source code of the proposed R-MCA methodology is available at http://data.snu.ac.kr/rmca.

INDEX TERMS Nearest neighbor, correlation, maximin, SOCP, QCLP, QP, regularization, kernel trick.

I. INTRODUCTION

Nearest neighbor (NN) classifiers [1], [2] are non-parametric methods that classify an object based on its distance to the nearest trained class. Owing largely to their simplicity and reasonable performance in practical problems, they have been widely used for various tasks such as image retrieval [3], video hashing [4], hyperspectral classification [5], query verification on road networks [6], and single particle tracking [7].

The main problems that arise with NN classifiers are that (1) it becomes computationally intensive to find the neighbors as the number of training samples increases and (2) the notion of nearest neighbors can break down in high-dimensional spaces. Approaches have been proposed to reduce the computation [8], [9] and to adaptively determine nearest neighbors (even in high-dimensional spaces) [2]. Template matching is another widely used technique that pre-computes a representative vector for each class and uses it to locate the nearest neighbor of an object [10]. In multiple subclass classification problems, where each class consists of multiple subclasses, a template is constructed for each subclass, and then the *aggregate* template of a class is created based on the subclass templates [11].

In this paper, we consider constructing the aggregate template based on the idea of the *maximin correlation approach* (MCA) [11]. For correlation-based NN classification problems, it is known that MCA can provide an optimal aggregate template in that MCA iteratively maximizes the minimum correlation with the templates it represents, eventually minimizing the maximum misclassification risk. MCA was originally proposed for multi-font optical char-

2169-3536 © 2016 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

acter recognition and has been successfully applied to automated protein function prediction [12] and typography clustering [13].

Despite the theoretical advantages of MCA, it has inherent limitations that have hindered wider applications in practice, such as susceptibility to noise and outliers, inability to handle nonlinearities in datasets, as well as high computational complexity. This paper proposes the *regularized maximin correlation approach* (R-MCA), a significantly improved solution method that overcomes these limitations of the original MCA. Regulariztion refers to techniques for alleviate the problem of overfitting, which may lead to unsatisfactory generalization performance of a learned model on new samples as the model is overly fitted to training samples. For example, the soft-margin *support vector machine* (SVM) [14] employs slack variables to balance the approximation and generalization capabilities of the model.

As opposed to the iterative method employed by the original MCA, we reformulate it as an instance of *quadratically constrained linear programming* (QCLP) [15]. The worst-case complexity of the iterative method grows quadratically as the number of objects increases. In contrast, the proposed QCLP formulation can be solved with linear complexity by the *interior-point methods* (IPMs) [16] when coefficient matrices are positive semidefinite.

Based on the QCLP formulation, we incorporate regularization and additional constraints that help R-MCA to find a robust representative vector even when (noisy) outliers exist. Our formulation has some resemblance to the regularization employed by the soft margin SVM [14]. We furthermore develop the Lagrangian dual of the regularized QCLP, which enables us to apply the kernel trick to effectively handle nonlinear structures possibly embedded in data. The kernel trick refers to the technique that replaces the inner products between objects in the input space by a *kernel function*, which allows us to map them to a (often infinite-dimensional) feature space and compute their inner products therein implicitly [17]. By this simple replacement, we can make linear algorithms learn nonlinear decision boundaries without explicit mappling. A well-known example is the kernel *principal component analysis* (PCA) [18], in which the linear operations of PCA [19] are carried out effectively in a feature space with an implicit nonlinear mapping by a kernel function.

This paper also presents experimental results that confirm the effectiveness of R-MCA on various public data sets. According to these results, the proposed R-MCA successfully delivers the following improvements:

- QCLP-based reformulation of MCA that enables acceleration, regularization and kernelization
- Regularization to fight overfitting and outliers
- Kernelization for discovering nonlinear structures

Note that R-MCA can contribute to devising a robust and scalable solution to not only nearest-neighbor classification but also a variety of other tasks based on finding group representatives. For such tasks, R-MCA can provide an alternative to conventional aggregates, such as centroids and medoids.

II. MAXIMIN CORRELATION APPROACH (MCA)

To make this paper self-contained, we briefly introduce the mathematical formulation of the MCA to the reader. Additional details can be found in [11].

Consider two non-zero vectors $\mathbf{u}, \mathbf{x} \in \mathbb{R}^m$. When $\mathbf{u}$ and $\mathbf{x}$ are column vectors, the centered correlation is defined as $\phi(\mathbf{u}, \mathbf{x}) = \mathbf{u}^T\mathbf{x}/||\mathbf{u}||_2||\mathbf{x}||_2$. MCA involves maximizing the objective function that is to find the worst-case value among the centered correlation between a non-zero vector $\mathbf{u}$ and all of the vectors in a set $\mathcal{X} \subseteq \mathbb{R}^m$. MCA can construct a template vector $\mathbf{u}$ that maximizes the minimum correlation by the following formulation:

$$\boxed{\begin{array}{l}\text{maximize } \min_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{u}, \mathbf{x}) \\ \text{subject to } ||\mathbf{u}||_2 \neq 0.\end{array}} \tag{1}$$

The optimization (1) is referred to as the *MCA problem* (MCAP). The original MCA [11] assumes that all of the $\mathbf{x}_i$'s are linearly independent, $||\mathbf{x}_i||_2 = 1$ for all $\mathbf{x}_i \in \mathcal{X}$, and $\mathbf{x}_i^T\mathbf{x}_j \geq 0$ for all $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$ (note that these assumptions are not required in the proposed R-MCA). An iterative solution to the MCAP was proposed in [11]: the template vector $\mathbf{u}$ is initialized to the centroid vector and is updated at each iteration to find the optimal vector $\mathbf{u}^\star$. For fixed m (the dimensionality), the worst-case complexity of this iterative algorithm is $O(n^2)$, where n is the number of objects in $\mathcal{X}$.

III. PROPOSED R-MCA METHDOLOGY

This section presents the details of the proposed R-MCA method. To propose more efficient solutions to the MCAP (1), we first formulate it as an instance of QCLP [15]. The QCLP formulation (2) enables us to find a solution using the general IPMs [16], instead of the iterative method proposed in [11]. The QCLP formulation also allows us to define slack variables that lead to a regularized version (3) to effectively handle outliers. From the regularized version (3), we further derive its Lagrangian dual form (7), which reveals the structure suitable for applying the kernel trick. To handle nonlinearities, we finally kernelize the dual form (7) into the kernelized R-MCA formulation (9).

Note that the original MCA (*i.e.*, the version without regularization) can also be kernelized; starting from the QCLP formulation (2), we derive its dual form (10) and the kernelized MCA (11).

A. GEOMETRIC INTERPRETATION

Fig. 1 shows the geometric interpretation and comparison of MCA and the proposed R-MCA, which will be formally defined in the next section. As shown in Fig. 1(a), solving MCA is equivalent to finding a template vector whose direction minimizes the worst-case angle between the vector and class members. With no outliers, the maximin template that MCA returns represents the group reasonably.

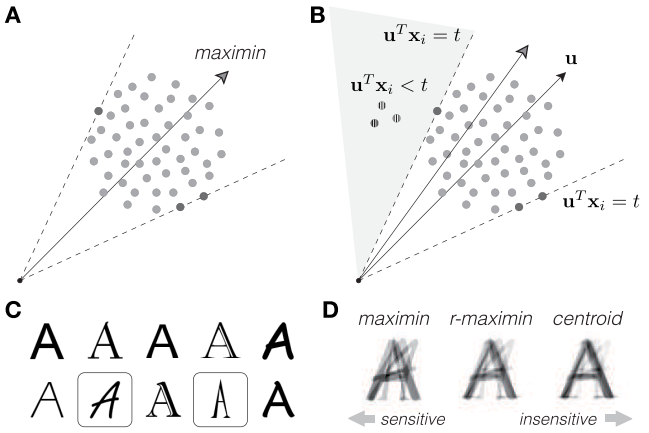


FIGURE 1. Geometric interpretation of regularization. (a) MCA finds a vector whose direction minimizes the worst (*i.e.*, maximum) angle between the vector and the class members. (b) Adding outliers (the shaded region) causes an abrupt swing in the traditional maximin that MCA returns. In contrast, the r-maximin that R-MCA finds is more robust to outliers. The objects on the dotted line from the origin have the minimum correlation with the template vector u. (c) The character 'A' represented in 10 different fonts. (d) The three aggregate templates of (c).

The existence of outliers significantly degrades the performance of MCA. For instance, Fig. 1(b) shows the scenario in which outliers are added to the data shown in Fig. 1(a). The maximin template returned by the original MCA swings abruptly towards the outliers because MCA does not recognize outliers. In contrast, the r-maximin template returned by R-MCA takes into account the outliers, yielding a template that represents the group more reasonably.

As an example from real applications, Fig. 1(c) shows the images of the character 'A' in ten different fonts and three types of templates, each of which aims at representing the ten images as a whole. In the centroid template, the two 'outlier' (boxed) fonts are averaged out and do not appear well, whereas the maximin template preserves them to some extent. For this reason, in multi-font character recognition, the maximin template, which incorporates outlier information, results in higher accuracy than the centroid template [11], [12]. In other applications, however, representing outliers may hurt classification accuracy. In R-MCA, we can adjust the sensitivity to outliers, providing an intermediate representation between the maximin and centroid templates (*e.g.*, compare the three templates in Fig. 1(c)).

B. QCLP FORMULATION OF MCA

A simple trick allows us to reformulate (1) as a tractable convex problem. After normalization of input vectors, (1) becomes equivalent to

$$\begin{aligned} &\text{maximize} \quad \min_{i=1,\ldots,n} (\mathbf{u}^T\mathbf{x}_i) \\ &\text{subject to} \quad ||\mathbf{u}||_2 \leq 1. \end{aligned}$$

The maximizer of the above maximin problem coincides with the solution of the following optimization problem:

$$\begin{aligned} &\text{maximize } t \in \mathbb{R} \\ &\text{subject to } \mathbf{u}^T\mathbf{x}_i \geq t, \quad i = 1, \ldots, n \\ &\qquad\qquad \mathbf{u}^T\mathbf{u} \leq 1. \end{aligned} \tag{2}$$

The equivalent formulation (2) for the MCAP with a finite set $\mathcal{X}$ is simple; it involves minimizing a linear function over $m + 1$ variables, with n linear equality constraints and one quadratic constraint. It is an instance of QCLP, a special type of optimization problem that can be solved globally and efficiently by the IPMs [16].

C. REGULARIZATION AND KERNELIZATION OF MCA

To construct a representative vector that is more robust to outliers (see Fig. 1(b) for an example), we apply the regularization to MCA. Regularization is a popular technique to prevent overfitting. Bertsimas and Copenhaver recently described a unifying view of the connection between robustification[1] and regularization [20].

Specifically, we introduce a non-negative 'slack' variable ξ_i for each object x_i, which can help the optimization problem find a solution insensitive to outliers. Using the slack variables, we can describe the regularized version of QCLP (2) as

$$\begin{aligned} &\text{maximize } t - \tfrac{\lambda}{n}\sum_{i=1}^{n} \xi_i \\ &\text{subject to } \mathbf{u}^T\mathbf{x}_i \geq t - \xi_i, \quad i = 1, \ldots, n \\ &\qquad\qquad \xi_i \geq 0, \quad i = 1, \ldots, n \\ &\qquad\qquad \mathbf{u}^T\mathbf{u} \leq 1 \end{aligned} \tag{3}$$

where λ is a user-specified sensitivity parameter for slack variables that serves as a regularization parameter; larger λ leads to a template vector that is more sensitive to outliers. Subsection III-E presents more details of λ and its effect on the solution of the optimization problem. Fig. 1 presents the geometric interpretation.

This formulation is similar to the optimization problem within the soft-margin support vector machine (SVM) [14], which is a relaxation of the original SVM. Leveraged by the regularization, the soft-margin SVM is more robust to labeling error, and we expect the proposed R-MCA to have the same advantage over the original MCA.

In order to facilitate understanding of (3), we derive its Lagrange dual [21] problem. First of all, we define the *Lagrangian* L: $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \mapsto \mathbb{R}$ associated with the problem (3) as

$$\begin{aligned} L(t, \xi, \mathbf{u}, \mathbf{v}, \mathbf{w}, z) = &-t + \frac{\lambda}{n}\sum_{i=1}^{n} \xi_i + z(1 - \mathbf{u}^T\mathbf{u}) \\ &- \sum_{i=1}^{n} \Big(v_i(\mathbf{u}^T\mathbf{x}_i - t + \xi_i) + w_i\xi_i \Big) \end{aligned} \tag{4}$$

[1] Immunizing a statistical problem against noise in the data.

where $\mathbf{v} = (v_1, \ldots, v_n)^T, \mathbf{w} = (w_1, \ldots, w_n)^T \in \mathbb{R}^n$, and $z \in \mathbb{R}$ are the Lagrange multipliers for the three inequality constraints of (3). We then define the *Lagrange dual function* g as the minimum value of the Lagrangian over t, ξ, and $\mathbf{u}$:

$$g(\mathbf{v}, \mathbf{w}, z) = \inf_{t,\xi,\mathbf{u}} L(t, \xi, \mathbf{u}, \mathbf{v}, \mathbf{w}, z). \tag{5}$$

To calculate the infimum of the Lagrangian, we partially differentiate the Lagrangian as follows:

$$\begin{aligned}
\frac{\partial L}{\partial t} &= -1 + \sum_{i=1}^{n} v_i = 0 \Rightarrow \sum_{i=1}^{n} v_i = 1 \\
\frac{\partial L}{\partial \xi_i} &= \frac{\lambda}{n} - v_i - w_i = 0 \Rightarrow \frac{\lambda}{n} = v_i + w_i \\
\frac{\partial L}{\partial \mathbf{u}} &= -\sum_{i=1}^{n} v_i \mathbf{x}_i + 2z\mathbf{u} = 0 \Rightarrow \mathbf{u} = \frac{1}{2z} \sum_{i=1}^{n} v_i \mathbf{x}_i.
\end{aligned}$$

From the above equalities, we can rewrite the Lagrange dual function (5) as $g(\mathbf{v}, \mathbf{w}, z) = -\frac{1}{4z}\mathbf{v}^T C \mathbf{v} - z$ where $C_{ij} = \mathbf{x}_i^T \mathbf{x}_j$. We can consider this as a function of z that is minimized when $z^\star = \sqrt{\mathbf{v}^T C \mathbf{v}}/2 \geq 0$. Thus, we can obtain a simplified representation of $g(\mathbf{v}, \mathbf{w}, z)$ as $-\sqrt{\mathbf{v}^T C \mathbf{v}}$ by substituting $z = z^\star$ into the above dual function and can finally formulate the dual problem of (3) as

$$\begin{aligned}
&\text{minimize } \mathbf{v}^T C \mathbf{v} \\
&\text{subject to } v_i \geq 0,\ w_i \geq 0, \quad i = 1, \ldots, n \\
&\qquad \lambda/n = v_i + w_i, \quad i = 1, \ldots, n \\
&\qquad \sum_{i=1}^{n} v_i = 1.
\end{aligned} \tag{6}$$

Additionally, we can combine the top two constraints of (6) into an inequality '$\lambda/n \geq v_i \geq 0$ for all i', because v_i and w_i are complements to each other. The problem now can be described as follows:

$$\boxed{\begin{aligned}
&\text{minimize } \mathbf{v}^T C \mathbf{v} \\
&\text{subject to } \lambda/n \succeq \mathbf{v} \succeq 0 \\
&\qquad \mathbf{1}^T \mathbf{v} = 1.
\end{aligned}} \tag{7}$$

We can identify that (7) is a convex quadratic program (QP) since the gram matrix C is positive semidefinite. Hence, when (7) has a feasible solution, by the strong duality principle [21], the template vector of R-MCA, $\mathbf{u}^\star \in \mathbb{R}^m$ (the primal solution), can be obtained from the solution of (7), $v^\star \in \mathbb{R}^n$ (the dual solution), as follows:

$$\mathbf{u}^\star = c^\star \sum_{i=1}^{n} v_i^\star \mathbf{x}_i \tag{8}$$

in which $c^\star = 1/\sqrt{\mathbf{v}^{\star T} C \mathbf{v}^\star}$.

D. KERNELIZATION

The nonlinearities in input space can often be handled better in high (possibly infinite) dimensional space. The mapping to and the computation in such high-dimensional spaces can be costly, if not impossible, but when the input data are used only through inner products, we can use the so-called *kernel trick* to perform implicit mapping and efficient computation.

Inspecting the dual form of (7) immediately suggests that we can apply the kernel trick to R-MCA. Replacing the inner products in (7) with a kernel matrix K yields

$$\boxed{\begin{aligned}
&\text{minimize } \mathbf{v}^T K \mathbf{v} \\
&\text{subject to } \lambda/n \succeq \mathbf{v} \succeq 0 \\
&\qquad \mathbf{1}^T \mathbf{v} = 1,
\end{aligned}} \tag{9}$$

where $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ for a Mercer kernel k. The kernelization allows the proposed R-MCA to find template vectors from data with nonlinearities, thus extending the applicability of the R-MCA. Section IV-D presents more details of the kernelization and supporting experimental results.

Similarly as in kernelized R-MCA, in order to obtain a dual for MCA, we first write Lagrangian of (2) and its partial derivatives as follows:

$$\begin{aligned}
L(t, \mathbf{u}, \mathbf{v}, z) &= -t - \sum_{i=1}^{n} v_i(\mathbf{u}^T \mathbf{x}_i - t) - z(1 - \mathbf{u}^T \mathbf{u}) \\
\frac{\partial L}{\partial t} &= -1 + \sum_{i=1}^{n} v_i = 0 \\
\frac{\partial L}{\partial \mathbf{u}} &= -\sum_{i=1}^{n} v_i \mathbf{x}_i + 2z\mathbf{u} = 0.
\end{aligned}$$

The Lagrange dual function of MCA thus becomes $g(\mathbf{v}, z) = -\frac{1}{4z}\mathbf{v}^T C \mathbf{v} - z$ just as in R-MCA. By inserting $z^* = \sqrt{\mathbf{v}^T C \mathbf{v}}/2$ into this Lagrange dual function, we can formulate the dual of the original MCAP formulation (2) as

$$\boxed{\begin{aligned}
&\text{minimize } \mathbf{v}^T C \mathbf{v} \\
&\text{subject to } \mathbf{v} \succeq 0 \\
&\qquad \mathbf{1}^T \mathbf{v} = 1.
\end{aligned}} \tag{10}$$

The above is the same as (7), except that the constraint $\lambda/n \succeq \mathbf{v}$ is missing. In other words, the dual of R-MCA (7) becomes the dual of MCA (10) if $\lambda > n$, hence the upper bound constraints in (7) disappears.

The kernelized version of the original MCA can also be derived in a similar way to Section III-C by replacing the dot-products in the dual quadratic program (10) with a kernel:

$$\boxed{\begin{aligned}
&\text{minimize } \mathbf{v}^T K \mathbf{v} \\
&\text{subject to } \mathbf{v} \succeq 0 \\
&\qquad \mathbf{1}^T \mathbf{v} = 1
\end{aligned}} \tag{11}$$

where the constraint $\lambda/n \succeq \mathbf{v}$ included in the regularized version (9) no longer appears.

E. ANALYSIS OF λ AND A COMPARISON WITH MCA

We elaborate on the characteristics of the template vectors obtained by R-MCA using the dual form (7). To satisfy the constraint $\mathbf{1}^T\mathbf{v} = 1$ therein, we consider the following four cases:

1) [$\lambda < 1$] If the Lagrangian multipliers v_i's are lower than $1/n$, the constraint $\sum v_i = 1$ cannot be satisfied. That is, (7) is not feasible if $\lambda < 1$.
2) [$\lambda = 1$] Because λ is the upper bound of v_i's, the only solution to fit the constraint $\sum v_i = 1$ must be $v_i = 1/n$ for all i. In this case, $\mathbf{u}^\star$ points to the same direction as the centroid of $\mathbf{x}_i$'s with the scaling factor ($\mathbf{u}^\star = c^\star \sum v_i^\star \mathbf{x}_i = c^\star \sum \mathbf{x}_i/n$).
3) [$1 < \lambda < n$] Larger λ makes v_i's less constrained; when λ becomes large, the upper bound constraints for v_i's become less restrictive for minimizing the objective $\mathbf{v}^T C\mathbf{v}$. Hence, the effect of each individual example $\mathbf{x}_i$, including the outlier, to the primal solution (8) can increase as λ increases.
4) [$\lambda \geq n$] If $\lambda \geq n$, the upper bound constraints for v_i's disappear; this follows from the fact that $\mathbf{v}$ is forced to be a probability vector by the other constraints, and thus it will always satisfy the upper bound constraints when $\lambda \geq n$. By comparing (7) with the dual of the original MCA formulation (2) as below, we deduce that the solution of R-MCA for this case coincides with that of MCA.

F. COMPLEXITY ANALYSIS

Recall that n corresponds to the number of objects and m corresponds to the dimensionality. Since the number of iterations that is necessary for IPM to find a solution is practically constant (typically from 10 to 50) [15], we can see that the QCLP (2) can be solved in $O(nm^2 + m^3)$ flops. For comparison, the number of flops required for the iterative method [11] is either $4mnp - mp^2$ or $4n^2p - 2np^2 + mn^2$, depending on the implementation, where p is the number of iterations. The empirical study in [11] shows that p grows nearly linearly in n. In result, MCA has order of $O(n^2m)$ or $O(n^3 + mn^2)$ time complexity due to the $p = O(n)$, while the proposed QCLP formulation takes $\min(O(nm^2 + m^3), O(n^3))$. The computational efficiency will be demonstrated in Section IV-E.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

We tested the proposed R-MCA methodology using the datasets listed in Table 1. More details about each dataset will be provided in the following subsections.

For our experiments, we implemented the proposed QCLP-based MCA and R-MCA solvers using SeDuMi software, a MATLAB toolbox for optimization over symmetric cones [26]. For comparison, we also prepared implementations of the original iterative solution to MCA as described in [11], the support vector machine (SVM), and the logistic regression.

TABLE 1. Data used in our experiments.

Name	n	m	Number of classes (description)
KSC [22]	5211	176	13 (land cover types)
MNIST [23]	10000	784	10 (digits '0'–'9')
SONAR [24]	208	60	2 (rock or mine)
GEO [25]	606	30954	2 (ulcerative colitis patient or not)
3D-NUT	272	3	2 (core or shell)

A. EFFECT OF REGULARIZATION ON SUBTYPE CORRELATION

To see the effect of regularization, we ran R-MCA with different values of parameter λ on a multiple-subclass dataset and measured the variation of correlation between subclass objects and the aggregate template. We used the Kennedy Space Center (KSC) dataset [22], which contains 5211 vectors with 176 dimensions. Each vector represents the signal intensities of different wavelengths measured above 13 types of land covers (105–927 vectors per class). Based on the characterization of vegetation, these classes can be grouped into three types or 'superclasses' (upland with seven land-cover subclasses, wetland with five, and water type with one).

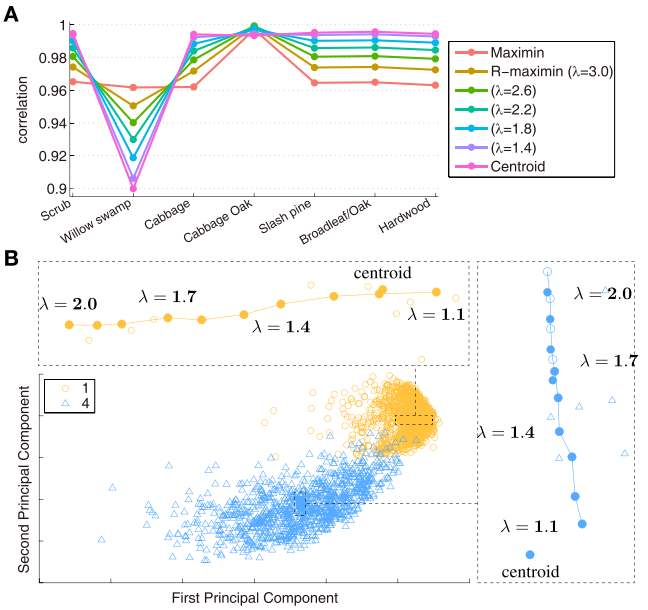


FIGURE 2. Effects of regularization and its parameter λ. (a) The minimum correlation between aggregate templates and subclass templates of the KSC data. (b) The part of the MNIST data representing the digits '1' (yellow) and '4' (blue).

Fig. 2(a) shows the correlation of seven subclasses of the 'upland' class with the regularized maximin aggregate templates (r-maximin) of five different λ values (1.4, 1.8, 2.2, 2.6, 3.0). As mentioned earlier, we define λ to manipulate the degree of regularization and can increase the best-case correlation value with the class members instead of sacrificing

the worst-case correlation. To verify this effect, the curves for the non-regularized maximin and the centroid template are also presented. As expected, the curves for the r-maximin are placed between the centroid and the non-regularized maximin.

B. EFFECT OF REGULARIZATION PARAMETER λ

In Section IV-A, we discussed that the regularization parameter λ works as a control knob that places the result from using the r-maximin somewhere between those from the centroid template and the non-regularized maximin template. To visualize the effects of varying λ, we utilized the MNIST database of handwritten digits [23]. From this database, we sampled 1135 and 982 images representing the digits '1' and '4', respectively. Each sample is a 28 × 28 image that can be represented by a 784-dimensional vector. We carried out the PCA of these samples and took the first two principal components only, transforming each of them into a 2-dimensional point, as shown in Fig. 2(b). In the figure, each of the two inlets magnifies the centroid and the r-maximin along with the corresponding images for visual inspection.

Recall from Section III-E that R-MCA eventually produces a centroid when $\lambda = 1$. As depicted in Fig. 2(b), we tested 10 different λ values of the interval [1.1, 2.0] to draw the trajectories of the r-maximin. The centroid in the class '1' is located in the upper-right region, because most samples in '1' class are distributed in that region. However, it is necessary to shift the aggregate template toward the outliers in order to minimize worst-case classification risk. We confirmed that reducing λ puts the regularized maximin template near the centroid, and increasing λ yields the r-maximin close to the outliers.

C. EFFECT OF REGULARIZATION ON CLASSIFICATION

To see the regularization effects in the context of classification, we carried out binary classification of the SONAR data [24], which consist of 111 mine-reflected and 97 rock-reflected sonar signals of 60 dimensions each. For NN classification using templates, we implemented the nearest template classifier that assigns an unknown vector to the class of its nearest (r-maximin, maximin, or centroid) template. For comparison, we also tested logistic regression, the linear SVM, and the RBF kernel SVM [27].

According to the experimental results from using neural networks in [24], nonlinearities exist in the distribution characteristics of the SONAR data. We thus preprocessed the data using the kernel PCA [18] with the Gaussian kernel ($\sigma = 1$). We then tested the five different classifiers with 2 and 5-fold cross-validation. The value of λ was determined by performing the CV with 4 different λ values (1.5, 2, 2.5, 3). The soft margin coefficient and the sigma of RBF kernel in SVM are 1 and 3, respectively.

Fig. 3(a) and (b) shows the receiver operating characteristic (ROC) curves from the first and second rounds of CV with $\lambda = 3$. The average area under the curve (AUC) values are 0.94, 0.90, 0.86, 0.89, 0.88, and 0.84 for NN with the r-maximin templates, the RBF SVM, the linear SVM, logistic regression, NN with the maximin templates, and NN with the centroid templates, respectively. Fig. 3(c) also presents an average and a standard deviation of 5 runs. The r-maximin classifier achieved 3.3% higher AUC on average than the alternatives.

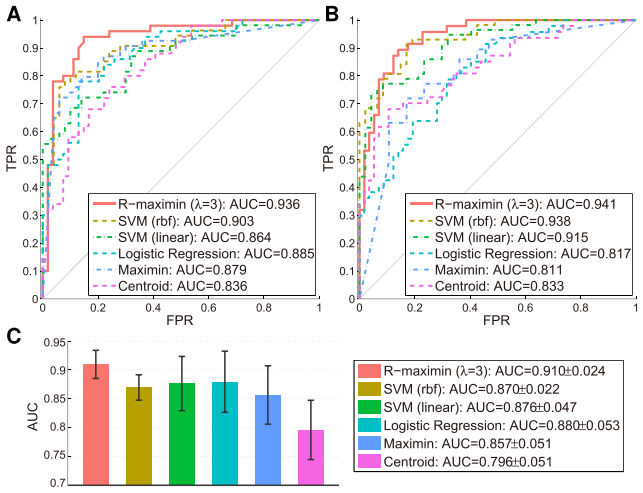


FIGURE 3. Regularization improves classification performance using SONAR. Hyper (a, b) ROC curves from 2-fold cross validation. (c) AUC values ($\mu \pm \sigma$) from 5-fold cross vaidation.

With respect to these AUC values, the r-maximin classification produced the best result, whereas the performance of the original maximin classification was lower than that of the SVM. This result suggests that the regularization can indeed improve the classification accuracy for real applications with noise.

D. EFFECT OF KERNELIZATION

Through kernelization, we expect R-MCA to become applicable to classification problems that contain complex shapes in the input space. Fig. 4 shows the result from a proof-of-concept experiment using a synthetic dataset termed 3D-NUT, which was generated as follows: we sampled a point $\mathbf{x} = [x_1, x_2, x_3]$ from a trivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, where $\mu = [0, 0, 0]$ and $\Sigma = \mathbf{I}$. For the sake of visualization, $\mathbf{x}$ was discarded if $x_2 < 0$ and $x_3 < 0$. Otherwise, we set the membership of $\mathbf{x}$ to the 'core' class if $||\mathbf{x}|| < 1$ and to the 'shell' class if $||\mathbf{x}|| > 2$.

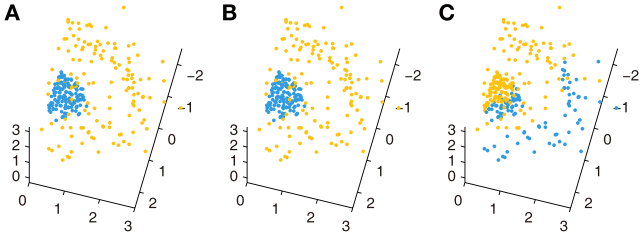


FIGURE 4. Effect of kernelization (data: 3D-NUT). (a) The true membership (best viewed in color). (b) The membership retrieved by the proposed kernelized R-MCA. (c) The membership assigned by the R-MCA.

Fig. 4(a) depicts the distribution of 272 points color-coded with binary membership (either 'core' or 'shell' class) in the input space. Applying the original R-MCA resulted in incorrect classification, as shown in Fig. 4(c). In contrast, the kernelized R-MCA (radial basis kernel with $\gamma = 1$) correctly separates the data points according to their membership, as shown in Fig. 4(b).

This experiment confirms that the kernelization works for R-MCA, and that we will be able to apply the kernelized version to other problems existing kernel-based methods (*e.g.*, kernel PCA) can be applied to.

E. COMPARISON OF EXECUTION TIME WITH MCA

We compare the runtime of the proposed QCLP-based solution and the original iterative solution [11] to the maximin correlation approach. To this end, we carried out two types of experiments. One is varying the number of objects n with the dimensionality m fixed, and the other involves varying m with n fixed. We measured the runtime using a Windows 7 PC equipped with an Intel i5-3570K CPU (3.4GHz, 6MB, 5GT/s) and 16GB RAM.

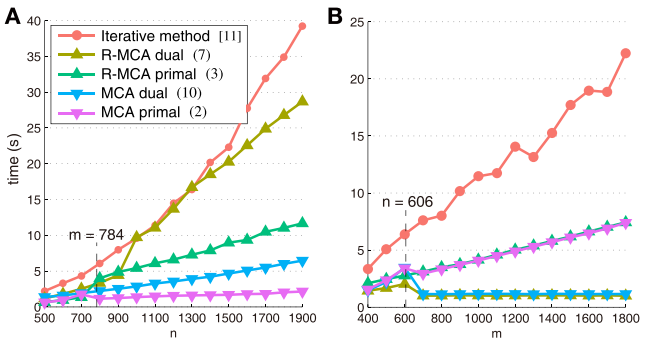


FIGURE 5. Comparison of execution times. Each time point represents the average of ten independent runs. (a) Varying n with fixed $m = 784$ (data: MNIST). (b) Varying m with fixed $n = 606$ (data: GEO).

Fig. 5(a) shows the varying-n fixed-m case for recognizing the digit '0' in the MNIST data (fixed $m = 784$). The time demand of the iterative solution remained the highest and also grew up faster than the others. As described in Section III, there are additional inequality constraints and variables in the regularized forms [(3) and (7)] in comparison with the original MCA [(2) and (10)]. Consequently, the two regularized versions require longer execution times than the unregularized ones when $n > m$, as shown in Fig. 5(a).

The varying-m fixed-n case is presented in Fig. 5(b). We used the NCBI GEO microarray dataset [28] (the accession number: GSE11223), which provides the regional variation of gene expression in ulcerative colitis patients [25]. The dataset has $m = 30954$ features and $n = 606$ samples (404 samples were generated by adding white Gaussian noise to the original 202 samples). Even though m increases, the runtime of the dual forms [(7) and (10)] does not increase noticeably, because $n \times n$ quadratic programming is involved in solving the dual forms. In contrast, the time demand of solving the primal forms [(2) and (3)] increases as m grows. Consequently, if $m > n$, the $n \times n$ quadratic programming would take less time, and solving the dual forms would be better.

Note that we can observe abrupt changes in runtime from both Fig. 5(a) and (b) at the point where $m = n$. This originates from the design of the SeDuMi toolbox. It uses an approximation based on the Farkas' lemma [21] and finds the solution $y \in \mathbb{R}^m$ such that $A^T y = 0$ if the solution $x \in \mathbb{R}^n$ does not exist for $Ax \geq 0$.

In summary, the primal and dual forms should yield the same solution, and we can always solve either the original MCA or the proposed R-MCA problems faster by using the proposed QCLP formulation than using the original iterative method. When $n > m$, using the primal forms [(2) and (3)] will be advantageous; otherwise using the dual forms [(7) and (10)] will be desirable. As the primal forms and the dual forms have $O(m)$ and $O(n)$ variables, respectively, the same observations can be made from the computational complexity of SeDuMi, which is $O(x^2y^{2.5} + y^{3.5})$ [26] (x is the number of variables, and y is the number of independent inequalities).

V. CONCLUSION

The maximin correlation approach (MCA) was originally proposed in the context of multiple-subclass classification problems that range from the optical character recognition problem to the automated protein family prediction. The aggregate templates found by MCA work well for such applications since they can minimize the maximum misclassification risk in the correlation-based nearest-neighbor classification setup. Nonetheless, practical limitations such as susceptibility to noise, inability to handle nonlinearities consideration, and high time demand have hindered a wider application of MCA to real applications.

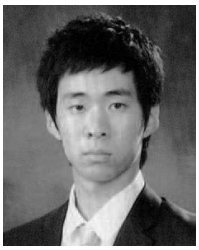
To address these drawbacks, we first described how to formulate the MCA as an instance of the QCLP and presented an efficient and general solution that can replace the original iterative solution. Based on this QCLP-based formulation, we further explained how to regularize and kernelize MCA in order to render it more robust to outliers and applicable to data with nonlinearities.

According to our experimental results, the proposed R-MCA successfully overcomes the limitations of the original MCA. Leveraged by the regularization, the proposed method outperformed the original MCA and the other alternatives tested in terms of classification performance. Given that the degree of regularization in R-MCA can be adjusted conveniently via a single parameter, the proposed R-MCA provides a flexible solution. In addition, we confirmed the computational benefit of the QCLP formulation and the effectiveness of kernelization in the (regularized) maximin correlation approach.

We anticipate that the kernelization and regularization of MCA will make MCA more appealing to a wider range of applications that we otherwise cannot satisfactorily analyze with the original MCA.

REFERENCES

[1] K. Fukunaga and T. E. Flick, "An optimal global nearest neighbor metric," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 3, pp. 314–318, May 1984.
[2] Y.-K. Noh, F. C. Park, and D. D. Lee, "Diffusion decision making for adaptive k-nearest neighbor classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1925–1933.
[3] S. McCann and D. G. Lowe, "Local Naive Bayes nearest neighbor for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3650–3656.
[4] G. Ye, D. Liu, J. Wang, and S.-F. Chang, "Large-scale video hashing via structure learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2272–2279.
[5] V. Menon, S. Prasad, and J. E. Fowler, "Hyperspectral classification using a composite kernel driven by nearest-neighbor spatial features," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 2100–2104.
[6] Y. Jing, L. Hu, W.-S. Ku, and C. Shahabi, "Authentication of k nearest neighbor query on road networks," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 6, pp. 1494–1506, Jun. 2014.
[7] J. Mazzaferri, J. Roy, S. Lefrancois, and S. Costantino, "Adaptive settings for the nearest-neighbor particle tracking algorithm," *Bioinformatics*, vol. 31, no. 8, pp. 1279–1285, 2015.
[8] Y. Tao and C. Sheng, "Fast nearest neighbor search with keywords," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 4, pp. 878–888, Apr. 2014.
[9] J. Wang, J. Wang, G. Zeng, Z. Tu, R. Gan, and S. Li, "Scalable k-NN graph construction for visual descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1106–1113.
[10] S. Korman, D. Reichman, G. Tsur, and S. Avidan, "FAsT-match: Fast affine template matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2331–2338.
[11] H. I. Avi-Itzhak, J. A. Van Mieghem, and L. Rub, "Multiple subclass pattern recognition: A maximin correlation approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 4, pp. 418–431, Apr. 1995.
[12] T. Lee, H. Min, S. J. Kim, and S. Yoon, "Application of maximin correlation analysis to classifying protein environments for function prediction," *Biochem. Biophys. Res. Commun.*, vol. 400, no. 2, pp. 219–224, 2010.
[13] T. Lee, S. J. Kim, E.-Y. Chung, and S. Yoon, "K-maximin clustering: A maximin correlation approach to partition-based clustering," *IEICE Electron. Exp.*, vol. 6, no. 17, pp. 1205–1211, 2009.
[14] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[15] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming," *Linear Algebra Appl.*, vol. 284, nos. 1–3, pp. 193–228, Nov. 1998.
[16] T. Terlaky, *Interior Point Methods of Mathematical Programming*, vol. 5. Norwell, MA, USA: Kluwer, 2013.
[17] M. A. Aizerman, E. M. Braverman, and L. I. Rozoner, "Theoretical foundations of the potential function method in pattern recognition learning," *Autom. Remote Control*, vol. 25, pp. 821–837, Jun. 1964.
[18] B. Schölkopf, A. J. Smola, and K.-R. Müller, "Kernel principal components analysis," in *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA, USA: MIT Press, 1999, pp. 327–352.
[19] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.
[20] D. Bertsimas and M. S. Copenhaver. (2014). "Characterization of the equivalence of robustification and regularization in linear, median, and matrix regression." [Online]. Available: http://arxiv.org/abs/1411.6160
[21] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[22] S. Kumar, J. Ghosh, and M. M. Crawford, "Hierarchical fusion of multiple classifiers for hyperspectral data analysis," *Pattern Anal. Appl.*, vol. 5, no. 2, pp. 210–220, Jun. 2002.
[23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[24] R. P. Gorman and T. J. Sejnowski, "Analysis of hidden units in a layered network trained to classify sonar targets," *Neural Netw.*, vol. 1, no. 1, pp. 75–89, 1988.
[25] C. L. Noble *et al.*, "Regional variation in gene expression in the healthy colon is dysregulated in ulcerative colitis," *Gut*, vol. 57, no. 10, pp. 1398–1405, 2008.
[26] J. F. Sturm, "Using SeDuMi 1.02, A MATLAB toolbox for optimization over symmetric cones," *Optim. Methods Softw.*, vol. 11, nos. 1–4, pp. 625–653, 1999.
[27] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi, "Kernel methods on the Riemannian manifold of symmetric positive definite matrices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 73–80.
[28] T. Barrett *et al.*, "NCBI GEO: Mining millions of expression profiles—Database and tools," *Nucleic Acids Res.*, vol. 33, no. 1, pp. D562–D566, 2005.



TAEHOON LEE received the B.S. and M.S. degrees in electrical engineering from Korea University, South Korea, in 2009 and 2012, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Seoul National University, South Korea. His current research interests include parallel computing, large-scale data analysis, and statistical learning.

TAESUP MOON (S'04–M'08) received the B.S. degree in electrical engineering from Seoul National University, Seoul, Korea, in 2002 and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 2004 and 2008, respectively. From 2008 to 2012, he was a Research Scientist with Yahoo! Labs, Sunnyvale, CA, USA, and he held a Post-Doctoral Researcher appointment with the Department of Statistics, College of Letters and Science, University of California Berkeley, Berkeley, CA, USA, from 2012 to 2013. From 2013 to 2015, he was a Research Staff Member with Samsung Advanced Institute of Technology, Samsung Electronics, Inc., Suwon, Korea. Currently, he is an Assistant Professor at the Department of Information and Communication Engineering, Daegu-Gyeongbuk Institute of Science and Technology (DGIST), Daegu, South Korea. His research interests include diverse areas such as statistical machine learning, information theory, signal processing, large-scale optimization, information retrieval, speech recognition, and remote sensing. He is a recipient of Samsung Scholarship.

SEUNG JEAN KIM (M'02) received the Ph.D. degree in electrical engineering from Seoul National University, Seoul, South Korea. Since 2002, he has been with the Information Systems Laboratory, Department of Electrical Engineering, Stanford University, Stanford, CA, where he was a Consulting Assistant Professor. He was then with Citi Capital Advisors, New York, NY. His research interests include convex optimization with engineering applications.

SUNGROH YOON (S'99–M'06–SM'11) received his B.S. degree in electrical engineering from Seoul National University, Korea, in 1996, and his M.S. and Ph.D. degrees in electrical engineering from Stanford University, CA, in 2002 and 2006, respectively. From 2006 to 2007, he was with Intel Corporation, Santa Clara, CA. Previously, he held research positions with Stanford University, CA, and Synopsys, Inc., Mountain View, CA. He was an Assistant Professor with the School of Electrical Engineering, Korea University, from 2007 to 2012. Currently, he is an Associate Professor with the Department of Electrical and Computer Engineering, Seoul National University, Seoul, Korea. His research interests include machine learning, data-driven artificial intelligence, and high-performance bioinformatics. He was a recipient of the 2013 IEEE/IEEK Joint Award for Young IT Engineers.