CHAPTER V

ON PREDICTION OPTIMIZATION METHODS

Multiple classifier systems are composed of the following components [52]: (1) *coverage optimization strategy*, that *generate* a mutually complementary classifiers that can be combined to achieve optimal accuracy, assuming that it has a fixed prediction combination function; (2) *prediction optimization strategy*, that inputs a fixed set of carefully designed and highly specialized classifiers, and we obtain a solution of an *optimal combination* of their decisions. It is also possible to apply the prediction optimization methods to classifiers generated with the aim of coverage optimization.

This chapter addresses the problem in prediction optimization methods, giving that the aim of coverage optimization method presented in Chapter 3 is satisfied. After a comprehensive introduction on several least square methods, we introduce the gradient descent approach to tune the parameter of ridge regressor, which is more reliable and computation attractive than traditional methods in prediction optimization of multiple classifier systems. Ensemble selection criteria in the form of diversity is also considered for improving the classification accuracy in a way that we can precondition the prediction optimization matrix by letting only the good ensemble components to be computed.

5.1 Introduction

From Chapter 2, we have seen that the problem of prediction optimization in MCS can be considered as a variance reduction technique in Monte Carlo Methods, so called *antithetic-common variates*. However, it is difficult to justify in practice whether the type of variates (antithetic or common variates) between a pair of ensemble members is satisfied the optimal condition of antithetic-common variates or not. In case of simple majority voting, we always prefer antithetic variates, since the variance of the classification output of the MCS can be reduced. The situation becomes harder in case of optimal linear combining, since the sign of the combining weights can be either positive or negative. This way, both antithetic and common variates can be used as one of the good variance reduction choices as long as we can assign the proper combining weights to the variates. In the situations that there are too many variates (ensemble members) that we can not fully assign the proper combining weights, we might not prefer variates that can harm the variance reduction process. Particularly, the unwanted variate is sometime called *harmful collinearity* member.

One of the approaches to improving the collinearity problems is to prune the harmful collinearity members, which was first studied by Hashsem [78]. In particular, Hashem used the BKW collinearity diagnostics [82] to select the harmful collinearity members. Particularly,

the harmful collinearity detection method can be considered as one of the feature selection techniques. This is consistently with the recent definition, which it is considered that multiple classifier systems can also be regarded as feature extractors [83]. Thus, Hashem's proposed method is not the only approach that can be used to overcome the harmful collinearity problem. Recently, there are many fitness functions and ensemble selection methods that are applied for selecting a candidate classifier subset from the generating multiple classifiers [84–86].

Alternatively, further improvements on the harmful collinearity problems are in the directions of the principal components and ridge regression estimators. These methods are biased-regression method known in the statistical community for more than thirty years. In fact, they are commonly used in statistics to control bias-variance tradeoff in predictions. The main contribution on the classification accuracy improvement is come from the capabilities of these methods to suppress (prevent) insignificant ensemble components from contributing to prediction combining.

5.2 Prediction Optimization Methods

Five different combining methods are briefly discussed here. The majority and the antithetic regression methods are directly related to the antithetic and common variates in variance reduction methods. Simple least square method is a simple method suggested for MCS in Reference [78]. Since each individual member decision matrix is not of full column rank, the solution of simple least square method always becomes ill-condition. In fact, this method do not consider the correlations between ensemble members, while the correlation based least square method do. Note that correlation based regression method still exhibits few ill-condition effects. To compensate the remained ill-condition effects, the principal component method was used for the pseudo inverse operation. In particular, the computation of pseudo inverse is based on Single Value Decomposition (SVD), where any singular valued less than a tolerance are treated as zero. Finally, instead of using principal component method for ridge regression method, we use gradient ridge parameter estimation to overcome the remained ill-condition effects. The reason for proposing the ridge regression method with gradient parameter estimation is that the SVD computation is computational expensive. All five different combining methods are compared for 3-class SAR ATR problem.

5.2.1 Majority Method

The classifier ensemble uses the simple majority decisions to make the final classification output. It is always used as a baseline method, especially due to its effectiveness when all classifiers are independent to each other.

5.2.2 Traditional Least Square Methods

It is widely accepted that the optimal weight combination rule [78,87] is closely related to the standard linear regression model. In our assumption, the elements of our desired output vector $\mathbf{Y}$ should be linear functions of all outputs in the ensemble members, $\mathbf{X}$:

$$\mathbf{Y} = \mathbf{X}\Gamma. \tag{5.1}$$

In the training phase, *the training output responses* $\mathbf{D}$ and *the training output responses* $\mathbf{X}$ can be written in matrix form as

$$\mathbf{D} = \mathbf{X}\Gamma, \tag{5.2}$$

where $\mathbf{D}$ is an $n \times q$ matrix of the training output responses (n is the number of training samples, q is the number of classes), the $\mathbf{X}_i(\mathrm{i} = 1, ..., \mathrm{r})$ are $n \times q$ matrices of outputs of the i-th ensemble members, $\mathbf{X}$ is an $n \times qr$ matrix whose columns are the $\mathbf{X}_i$, the $\Gamma_i(\mathrm{i} = 1, ..., \mathrm{r})$ are $q \times q$ matrix of weight parameters of the i-th ensemble members, and

$$\Gamma = \begin{bmatrix} \Gamma_1 \\ \Gamma_2 \\ \vdots \\ \Gamma_r \end{bmatrix} \tag{5.3}$$

is an $qr \times q$ matrix whose rows are the Γ_i.

The optimal combination-weights estimated by least squares method can be obtained as

$$\hat{\Gamma} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{D}. \tag{5.4}$$

If the above model is assumed to have a constant term, one of the $\mathbf{X}_i$'s, usually the first, will be assumed to be the column vector of q ones. The inclusion of the constant term helps in correcting for (possible) biased in the ensemble members.

Literally, the term $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ is the pseudo inverse of $\mathbf{X}$, if $\mathbf{X}^T\mathbf{X}$ is nonsingular. In the case that $\mathbf{X}$ is not of full column rank, this solution becomes ill-condition. In that case this optimization problem can be solved by either using suboptimal methods [87], i.e., sequential approach or singular value decomposition approach, or use dummy augmentation to make X a full column rank in a higher dimensional space and then solve the problem. As proposed in Reference [88], the harmful collinearities can also be pruned out by the genetic-based algorithm. In addition, this optimization problem can also be solved using suboptimal methods [87], i.e., sequential approach or singular value decomposition approach. Another possibility of solving the ill-conditioned combination-weights is by detecting the presence of collinearities and pruning the *harmful collinearity* members [78].

In the events that ensemble members were often trained independently or sequentially, it seems appropriately to use one of the above approaches. Anyway, all of the above optimal weight combination schemes do not consider the correlations between ensemble members.

This is why the harmful collinearity selection approach is preferred to be used with the above weight combination schemes. However, there are several weaknesses of the pruning approaches [78] that make them less desirable to be used with the above weight combination schemes. First, the algorithms is not only greedy but also limited by the ability or the cost of acquiring extra data for validation. Second it is possible that the selection algorithm may allow dropping too many ensemble components before their diverse information (knowledge) will be accounted. In other words, dropping ensemble members often introduced biased to the combination weights estimation.

Part of this problem arises from its inefficient weight optimization method. To be more elaborated, the following stages for implementing the harmful collinearity selection algorithm will be described. The first stage is the weight optimization method, while the rest of the stages are related to the detection of the presence of collinearities and the harmful collinearity component pruning. Indeed, the normal regression equation presented in the first stage is based on the independency of the ensemble members. This is the major problem since the weights will be inaccurately optimized if the ensemble members are correlated. This way, the false detection of the harmful collinear ensemble components may be caused from this inaccurately optimization.

Recently, further improvements on the harmful collinearity problems are in the directions of the principal components and ridge regression estimators. Next, we will present several least square methods aimed at tackle the harmful collinearity problem, especially one of the early methods, called *antithetic regression* method, proposed to compensate the correlation problem.

5.2.3 Antithetic Regression Method

The idea of regression method of antithetic variates is first discussed in Reference [89]. It is based on making allowances for various causes of variation in the data (correlated output observations). The introduced correlations among the output observations arise because the observation outputs are influenced by certain concomitant conditions of the experiment (e.g., linear transformation or perhaps the multiple description transform coding approach in our case). When the influential effect of concomitant variables are observed and determined, we can then smooth out its effects from the output observations, thus leaving only those information which are not due to the concomitant conditions.

In the context of correlated observations, concomitant numbers are purposely used as the relative measures for characterizing correlation among experiments. Concomitant numbers can be valued 1 or 0 for the purpose of data classification into strata or categories. In this case, this method is known as the *analysis of variance*. In mixed case, when some but not all of concomitant numbers are restricted to the values 0 and 1. One has what is called the analysis of covariance. For the simplicity purpose, we follow the assumption used in Reference [89], where the concomitant numbers are known with values 0 or 1.

Instead of using linear model as (5.2), Hammersley and Handscomb [89] proposed that

a vector, whose elements are the estimation values of antithetic estimators, can be represented by reparameterization of unknown estimands with a concomitant vector (matrix). It is not interesting to see that we reuse the same cross-valiation data in (5.2) by substituting cross-validation output response D in the least squares equation at the training phase of the prediction optimization. A first step toward the improvement of (5.2) to harmful collinearity problems is by extending the basic least squares estimation to the case where there are correlation among outputs.

Let recall (5.2)

$$\mathbf{D} = \mathbf{X}\beta. \tag{5.5}$$

Solving (5.2) by the method of maximum likelihood [90, page 621] leads to optimal combination-weight matrix

$$\widehat{\beta}_{LS} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{V}_{\mathbf{D}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{V}_{\mathbf{D}}^{-1}\mathbf{D}, \tag{5.6}$$

where $\mathbf{V_D}$ is the covariance matrix of $\mathbf{D}$, where the size of $\mathbf{V_D}$ is $n \times n$.

Note that this regression method still exhibits few ill-condition effects. To compensate the remained ill-condition effects, the principal component method was used for the pseudo inverse operation. In particular, the computation of pseudo inverse is based on Single Value Decomposition (SVD), where any singular valued less than a tolerance are treated as zero.

5.2.4 Ridge Regression

Recently, the further improvements for the collinearity problems are in the directions of the principal components and ridge regression estimators (there is also some discussions [90] that ridge regression is equivalent to principal component estimator). In the ridge regression approach, instead of diagnosing which ensemble components are harmful, a further step toward the improvement of (5.2) is to introduce the biased parameter δ (or equivalently called *ridge parameter*) to the traditional regression algorithm, and use it to derive optimal weight combining rules. In this case, ensemble component selection and combination-weights estimation are performed simultaneously through ridge estimation.

The ridge estimator for the linear model, denoted by $\widehat{\Gamma}_R$, is defined by

$$\widehat{\Gamma}_R = \left(X^* V_D^{-1} X + \delta T\right)^{-1} X^* V_D^{-1} D, \tag{5.7}$$

where T is some positive definite matrix (very often T is chosen to be equal to identity matrix in practical applications).

Beyond the equivalency to the antithetic regression with principal component method, ridge estimator is also equivalent to a new least square method, so called the *covariance shaping least square* estimator, where it is interpreted and inspired from the framework of quantum signal processing [91]. With little manipulation, we can derive the covariance of the ridge estimate $\widehat{\Gamma}_R$ as

$$\widehat{R}_R = \left(I + \delta(X^* V_D^{-1} X)^{-1} T\right)^{-1} (X^* V_D^{-1} D + \delta T)^{-1}. \tag{5.8}$$

This way, when we decide to do the estimation based on the minimization of the (weighted) total error variance in the observations subject to a constraint defined by the above covariance, we can control the dynamic range and spectral shape of the covariance of the estimation error. This is a biased estimator directed at improving the performance of the traditional least squares estimator at low to moderate signal-to-noise (SNR). It should be noted that least squares estimation at low to moderate SNR can be viewed as the situation where base classifiers are weakly trained.

In most of related applications [92], the ridge parameter δ is chosen by cross- validation, in which some part of the training sample is held back and the value that best predicts this held-back data is our estimate. While there is nothing wrong with this grid search in low dimensions, getting the right scale for the parameter (or the interval of the grid search) might turn out to be a difficult task. The ridge regression literature discusses a plethora of methods to estimate δ from the training data. At least it is a good idea to use them first to get the correct scale for the parameter followed by cross-validation, especially for the merit of faster search (computations). This chapter discusses one of the methods and point out its relevant to harmful collinearity problem.

Moreover, it is possible to extend the optimal combining-weight method based on the concept of harmful collinearity suppression as in (5.6) by also constraining on the covariance of the estimation error. It is also possible to get the correct constraint covariance R as in (5.8)) and use it to solve with the covariance shaping least square estimator.

5.2.4.1 Gradient Ridge Parameter Estimation

The parameter δ is sometimes called the "*ridge parameter*". In fact, it is used for demonstrating on how much the least squares coefficient shrunk toward 0. As argued in Reference [92], the ridge parameter can be chosen by either choosing it a priori or estimating. One of the estimating methods is proposed by Hoerl, Kennard and Baldwin (HKB) [92]. If r is the number of ensemble components and n is the number of training samples, the HKB estimator is originally derived from

$$\delta = \frac{n\sigma^2}{\mathbf{\Gamma}^{*}\mathbf{\Gamma}}. \tag{5.9}$$

Let $\widehat{\mathbf{\Gamma}}$ denote the ordinary least squares estimator of $\mathbf{\Gamma}$ and

$$\widehat{\sigma}^2 = \frac{\left(\mathbf{D} - \mathbf{X}\widehat{\mathbf{\Gamma}}\right)^{T}\left(\mathbf{D} - \mathbf{X}\widehat{\mathbf{\Gamma}}\right)}{n - r - 1} \tag{5.10}$$

is the estimator of σ^2. Let $\widehat{\mathbf{\Gamma}}_R(\delta')$ denote the ridge estimate of $\mathbf{\Gamma}$ at $\delta = \delta'$. At the t^{th} iteration,

$$\delta_t = \frac{n\widehat{\sigma}^2}{\widehat{\mathbf{\Gamma}}_R^{*}(\delta_{t-1})\,\widehat{\mathbf{\Gamma}}_R(\delta_{t-1})}, \tag{5.11}$$

with $\delta_0 = \frac{n\hat{\sigma}^2}{\Gamma^* \Gamma}$. In fact, the new "sensible" ridge parameter δ_t is chosen such that the difference between successive δ is as small as possible. In this case, the criterion function for choosing an appropriate value of ridge parameter can be defined by

$$\rho = \delta_t - \delta_{t-1}, \tag{5.12}$$

often ρ is chosen to be equal to 10^{-4} in practical applications. In other words, ρ is used as the stopping criterion for the ridge parameter estimation.

Evidently, the parameters δ and ρ can be used indirectly to tackle harmful collinearities among all ensemble components. For example, if the true value of the ridge parameter is equal to 0, it is an indicator that all ensemble component are independent to each other, or an ordinary least square estimator is preferred. The larger the value of the ridge parameter, the more the collinearity components are suppressed. In this case, the predictors will not play any role at all in classification resulting in a total useless classifier. Instead of using the harmful collinearity identification criterion [78] to rule out the harmful collinearity components, ridge estimator tries to suppress ensemble components with serious collinearity.

5.2.4.2 Connection with Principal Component Method

The optimal weight combination rules of (5.7) and Equation (5.9)- (5.12) can be considered as one of the solutions of the linear model of (2.10). Specifically, let the $\widehat{\beta}_{R,m}\,(m = 1, ..., r)$ are $q \times q$ matrix of weight parameters of the m-th ensemble members, and

$$\widehat{\beta}_R = \begin{bmatrix} \widehat{\beta}_{R,1} \\ \widehat{\beta}_{R,2} \\ \vdots \\ \widehat{\beta}_{R,r} \end{bmatrix} \tag{5.13}$$

is an $qr \times q$ matrix whose rows are the $\widehat{\beta}_{R,m}$, and the $\mathbf{X}_\mathbf{i}'(\mathbf{i} = \mathbf{1}, ..., \mathbf{r})$ are q-dimensional vectors of outputs of the i-th ensemble members, $\mathbf{X}'$ is an $1 \times qr$ matrix whose columns are the $\mathbf{X}_\mathbf{i}'$, then with $\mathbf{c_m} = \widehat{\beta}_{\mathbf{R},\mathbf{m}}$, $f(\mathbf{x}, \mathbf{p_m}) = \mathbf{X}_\mathbf{i}'$, the final decision output function $\widehat{F}(\mathbf{x})$ will be equal to $\mathbf{X}'\widehat{\beta}_\mathbf{R}$.

To explore on how the role of ridge parameter suppresses the components with serious collinearity, suppose that we transform the ridge equation in (5.7) using the matrix of eigenvectors as in the principal component method. For the purpose of illustration, we can represent (5.7) in another matrix form by making an assumption that the matrix $\mathbf{V_D}$ in (5.7) is equal to identity matrix. Hence, (5.7) becomes

$$\left(\mathbf{X^T X} + \delta \mathbf{T}\right) \beta_\mathbf{R} = \mathbf{X^T D}. \tag{5.14}$$

Transforming the ridge equations in (5.14), we obtain

$$\left(\mathbf{\Lambda} + \delta \mathbf{T}\right) \gamma_\mathbf{R} = \mathbf{Z^T D}, \tag{5.15}$$

where $\gamma_{\mathbf{R}} = \mathbf{P}^{\mathbf{T}}\beta_{\mathbf{R}}$, $\mathbf{Z} = \mathbf{XP}$, $\mathbf{P}$ denotes the orthogonal matrix of eigenvectors of the correlation matrix $\mathbf{X}^{\mathbf{T}}\mathbf{X}$, and $\mathbf{Z}^{\mathbf{T}}\mathbf{Z} = \mathbf{\Lambda}$.

Then, we can obtain eigenvalues for the coefficient matrix in the ridge equations as $\lambda_i + \delta$. Thus, if $\lambda_{\min}$ indicates a serious collinearity, we simply add the constant δ so that this is no longer the case. It follows that the j^{th} component of $\gamma_{\mathbf{R}}$ is estimated by

$$\widehat{\gamma}_{Rj} = \frac{\lambda_j}{\lambda_j + \delta}\widehat{\gamma}_{\mathbf{j}}, \tag{5.16}$$

where $\widehat{\gamma}_{\mathbf{j}}$ is the least squares estimate of $\gamma_{\mathbf{j}}$. We can see that the factor $\frac{\lambda_j}{\lambda_j+\delta}$ will be close to 1, if λ_j is large relative to δ. On the other hand, when the value of λ_j is small relative to δ, this factor will be close to 0. In this case, $\widehat{\gamma}_{Rj}$ will also be closed to 0, which will be an indicator that the harmful collinearity components are suppressed.

5.3 Discussion on Prediction Optimization Methods

Recall that there are several weaknesses of the pruning approaches based on the harmful collinearity detection proposed in Reference [78]. This is from the reason that the ordinary least square method used in the method are very unstable for computing combining-weights, giving the harmful collinearity detection far from successes. Thus, one should look for some remedial solutions that is attractive in the sense that it will not depend on the ordinary least square estimates.

There are several remedial solutions for solving the harmful collinearities, which are applicable to both fix and optimal weight combination rules. The first remedial solutions for fixed combining rule, called *regression method* of *correlated variates*, was proposed by Hammersley and Handscomb [89, pages 19, 23, 66] and Aitken (see details in References [90, page 78], [93, page 221]). The method is a more advanced way to take advantage of several correlated ensemble components by eschewing the unconcerned correlation based combination rules to a concerned one. It should be noted that the method can be considered as the generalization of the fix combination rule (equal weight combination rule) to the heteroscedastic case, where the heteroscedastic situation is the situation that observations of members do not have equal covariance matrices.

To the best of our knowledge, we have learned that these regression (least squares) methods [89]- [94] are as untried as they are new to the ensemble learning community. The methods based on regression methods of correlated variates seem to be a highly promising means for examining the mechanism of ensemble learning, especially underlying the linear transformations of data. Recently, the advanced concept of combining method, called *feature based decision aggregation architecture* [95] is proposed in the MCS community. In fact, it can be considered as a variant method of system identifications by regression methods [96]. Consequently, any improvement in the least squares techniques [92, 97] can be of interest to the optimal combining-weights techniques.

Note that it might be too early to claim that this approach is better than other two-stage combination-weights estimation schemes. The schemes that exploit some forms of feature selection procedures at the output stage of the multiple classifiers before the prediction optimization. One of the explanations for recommending the use of ridge regression with estimated ridge parameter is inspired from arguments regarding to the classification performance between generative and discriminative classifiers. Over the arguments between generative and discriminative classifiers, there are several compelling reasons for using discriminative rather than generative classifiers, one of which, succinctly articulated by Vapnik [32] is that

> *"One should solve the problem (classification) directly and never solve a more general problem as an intermediate step (such as modeling posteriori probability)."*

At this point in the discussion, it seems to be reasonable to modify the Vapnik's statement (above) to

> *"One should solve the biased least squares estimation problem directly and never solve a more general problem as an intermediate step (such as diagnosing harmful collinearity or computing diversity)."*

Note that these regression methods aim at reducing the harmful collinearity problem, but their estimation still exhibits few ill-condition effects. To further compensate the remained ill-condition effects, the accuracy can be improved by promoting the diversity among the selected classifier members that tend to error in different subareas of the instance space. The main reason is that positively correlated classifiers only slightly reduce the added error, uncorrelated classifiers reduce the added error by a factor of $1/L$, and negatively correlated classifiers reduce the error even further. This is indeed the physical meaning of the antithetic and common variates of variance reduction technique in Monte Carlo methods.

In fact, several researcher considers multiple classifier systems as feature extractors [83] and the needs for ensemble selection [84–86]. Next, we discuss several fitness functions that can be used to compensate the remained ill-condition effects left from estimates. Here, the search strategy for ensemble selection is by ranking for the best group of classifier members. Moreover, more advanced strategies inspired from general feature selection techniques [84, 85] can also be applied to ensemble selection.

5.4 Diversity Measures

There are several statistics to assess the similarity of two classifier outputs. These statistics can be derived from the ratios between various quantities of occurrences of the correct/incorrect types of outputs. The pairwise diversity measures of the similarity of two

classifier outputs used in this work are borrowed from Reference [86], and shown as follows.

The Q Statistics

The level and sign of dependency between a pair of classifiers with binary outputs are defined by

$$Q_{i,j} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}, \tag{5.17}$$

where N^{11} is the number of occurrences both classifiers are correct, N^{00} is the number of occurrences both classifiers are incorrect, and N^{01} and N^{10} the number of occurrences when both classifiers make different decision and either one of them is incorrect.

Note that all measures listed below are the pairwise diversity measures. For set of more than two classifiers, the mean value of the pairwise measure is considered to be the measure value for that set, that is

$$Q_{av} = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} Q_{i,j}, \tag{5.18}$$

here, L is the total number of predictors used in the ensemble. The smallest measure value will indicate the best subset of member classifiers.

The Correlation measure ρ

The correlation between two binary classifier outputs using occurrence frequency of the quantity measures as in the Q Statistics derivation, or

$$\rho_{i,j} = \frac{N^{11}N^{00} - N^{01}N^{10}}{\sqrt{(N^{11} + N^{10})(N^{01} + N^{00})(N^{11} + N^{01})(N^{10} + N^{00})}}. \tag{5.19}$$

The disagreement measure S

This measure is defined as the ratio between the number of occurrences on which both classifiers make different decision and either one of them is incorrect to the total number of decisions. This can be

$$S_{i,j} = \frac{N^{01} + N^{10}}{N^{11} + N^{00} + N^{01} + N^{10}}. \tag{5.20}$$

The double-fault measure D

This measure is defined as the ratio between the number of occurrences when both classifiers make incorrect decisions. This can also be represented as

$$D_{i,j} = \frac{N^{00}}{N^{11} + N^{00} + N^{01} + N^{10}}. \tag{5.21}$$

5.5 Experimental Results

In this experiment setting, the available training data were divided into two parts. The first part is used for coverage optimization, and the second is used for prediction optimization. In particular, 1621 and 135 samples were randomly selected for the first and second parts, respectively. Here, we used one-fold cross-validation to obtain the least square estimates. We constructed 9 classifiers for the 3-class ATR problem by using the MC-LDB algorithms. Next, we combined the classifier outputs using the ridge estimation scheme with gradient ridge parameter estimation as the prediction optimization method and compared the recognition accuracy with the other methods.

As presented in Table 5.1, we observe that the performance of our prediction optimization method seems to be comparable with the majority and the principal component approaches. Our proposed method is better than the simple majority method at the small target window (32x32), and slightly less accurate than the simple averaging and the principal component approaches at the large target windows. One of the reasons for our moderate performance should lie on the fact that the estimated ridge parameter might be adjusted to 0 too aggressively. Thus, some of the informative decisions might be suppressed unintentionally by the algorithm. Another reason lies on the disadvantage of the weight combining scheme. Previous work [66] had shown that optimizing the combining weights can lead to overfitting while an unweighted voting scheme is generally resilient to the problems of overfitting. The third reason is that the prediction optimization should be avoid if the base classifiers are (almost) overtrained [98], which was the case in this experiment setting.

Practically, we can resolve this problem by reducing the number of samples used at the coverage optimization level and increasing the number of samples for combining-weights estimation at the prediction optimization level; or using the k cross-validation. It should be mentioned that one of the advantages of our proposed method over the principal component method is that the computation complexity in our proposed method is less expensive than the principal component method.

We previously experienced with the optimal weight combining rules without diversity measures. To compare the performance of optimal weight combining rules with diversity measures, we used diversity measures to select 7 best diverse classifiers from 9 classifiers. Our baseline weight combining rules were majority and other least square methods implemented on classifiers trained by 9 descriptions without the use of diversity measures. From our experiments with Q statistics, Figure 5.1 shows the high, low, and variance values of recognition accuracy evaluated over a range of numbers of coefficients per description at different window sizes. We can see that several methods show the robustness of weight combining rules in terms of numbers of coefficients per description. For example, majority rules with/without the use of Q statistics, principal component based antithetic regression with the use Q statistics, and ridge regression with/without the use of Q statistics, were

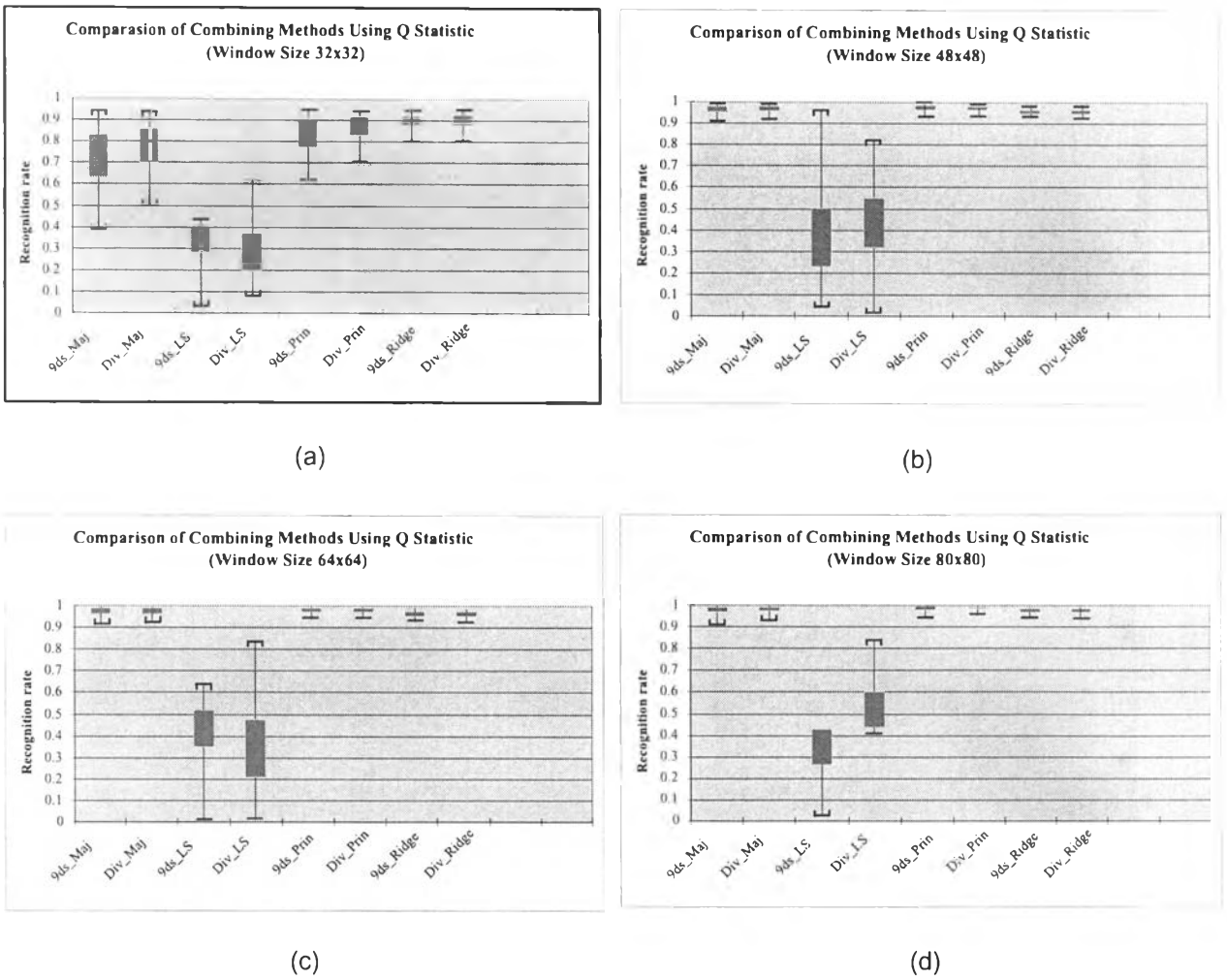


Figure 5.1: Comparison of different combining methods with Q diversity measure at various window sizes with the overall percentage normalization to unity. a) 32 x32, b) 48x48, c) 64x64, and d) 80x80. $9ds\_Maj$ means that all 9 descriptions are used with majority combining. Div_Maj means that 7 descriptions are selected by using Q statistics. After selection, they are integrated by majority combining. $LS, Prin$, and $Ridge$ are represented for traditional least square regressor, antithetic regressor with principal component approach, and ridge regressor with ridge parameter estimation, respectively. Note that each graph is plotted the high, low, upper-half standard deviation, and lower-half standard deviation of the recognition accuracy derived from various number of coefficients per descriptions.

among the methods that are robustness in terms of numbers of coefficients per description.

We summary several good weight combining rules based on their performance in Figure 5.2. From Figure 5.2(b), the principal component methods with/without the use of Q statistics outperform the majority methods with/without the use of Q statistics in terms of low and variance of recognition accuracy. As detailed in Table 5.2, we present the experimental results with respect to the window sizes and the optimal numbers of coefficients per description.

From the experimental results, we found that the best performance of the recognition accuracy is 99.71 percent, while most of the MCS methods presented here were able to achieve 99.63 percent, which is close to the best performance. In fact, the principal component approach with the use of disagreement measure slightly outperformed other methods (its highest recognition accuracy reached the best performance). One of the reasons for better performance of the principal component approach is that we used singular value decomposition to implement the pseudo inverse in (5.6). However, ridge regression is still a good alternative for computing optimal weight combining rule, since we made no use of the highly computation principal component method. Furthermore, we found that simple least square method were least stable as we expected. Thus, its uses for selecting the component neural networks proposed in Reference [78] should be performed with caution.

In summary, the assumption that the diversity measures are necessary for evaluating the potential candidates of the generating multiple classifier members is valid. For majority combining method, the reason that all classifier components of generated form MD-LDB are independent is partial correct, since there are still dependence left in the selected candidates for our MD-LDB. As a result, it is still required to do the computation for the optimal combining-weights, and this fact is supported by our experiments.

5.6 Conclusions

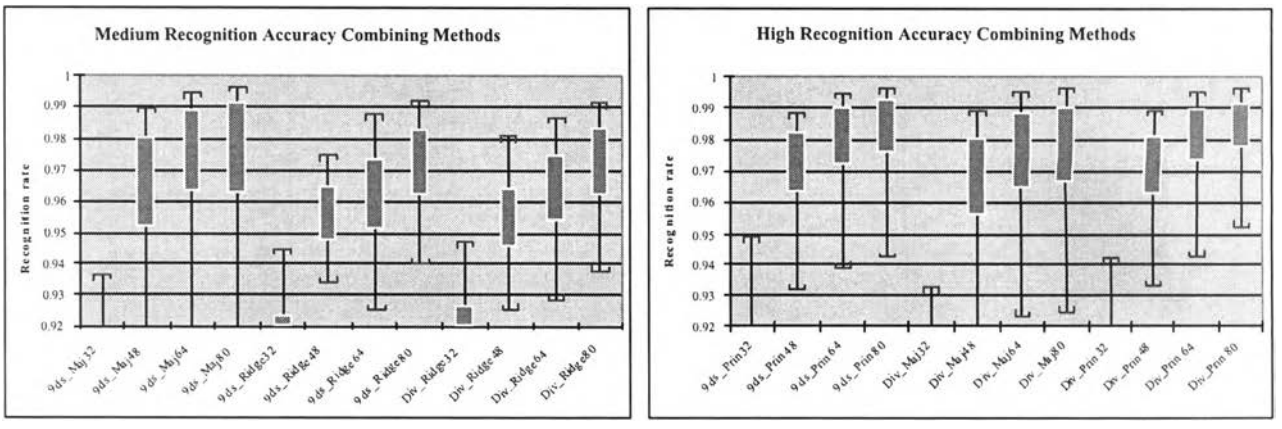
This chapter compares several least square estimation techniques and discusses the singularity of the ensemble output matrix that contributes to the ill-conditioned effect (or harmful collinearity problem) in multiple classifier systems. Inspiring from the early least square methods that proposed to overcome the correlated variates estimates, we study several least square methods that can be used to alleviate the harmful collinearity problem. In several methods, it is necessary to use ensemble selection for improving the accuracy.

The main results of this chapter can be summarized as follows:

- The majority, and principal component based ridge regression methods with/without diversity measures give the comparable best performance.
- The antithetic regressor exactly outperforms the ordinary least square estimators. Thus, the use of ordinary least square estimate in the original harmful collinearity detection should be performed with more precautions.
- The disagreement measure S consistently gives the best performance at the large window sizes.

Table 5.1: Comparison of different least square methods in overall percentage of images correctly recognized as a function of image sizes.

Methods / Image Size	32x32	48x48	64x64	80x80
MC-LDB with Simple Majority	84.69	98.53	99.34	99.49
MC-LDB with Principal Component	**94.95**	**98.83**	**99.49**	**99.63**
MC-LDB with Ridge Regression	94.51	97.44	98.75	99.19



(a) (b)

Figure 5.2: Comparison of various methods for medium and high recognition accuracy with the overall percentage normalization to unity. Note that each graph is plotted the high, low, upper-half standard deviation, and lower-half standard deviation of the recognition accuracy derived from various number of coefficients per descriptions. a) Medium recognition accuracy. b) High recognition accuracy. Medium recognition accuracy means the prediction optimization method that has its mean and standard of deviation of the recognition accuracy in the middle range in term of the number of coefficients per description. This is similar to high recognition accuracy as well.

Table 5.2: Comparison of different least square methods using various diversity measures. These are the best performances in overall percentage regarding to the optimal numbers of coefficients per description

Method	Window Size	Diversity measures				
		All	Q	ρ	S	F
Majority	32x32	93.63	93.26	93.26	94.21	93.26
	48x48	98.97	98.97	98.9	98.68	98.75
	64x64	99.41	99.56	99.34	99.41	99.34
	80x80	**99.63**	**99.63**	**99.63**	**99.71**	**99.63**
Simple LS	32x32	43.59	61.25	47.4	50.04	47.4
	48x48	95.6	81.98	84.4	80.73	75.09
	64x64	63.74	83.22	83.22	83.22	83.22
	80x80	43.0	83.08	83.08	83.08	83.08
Principal Component	32x32	73.19	81.1	81.1	83.15	81.1
	48x48	98.83	98.9	98.75	98.68	98.83
	64x64	99.49	99.56	99.34	99.49	99.34
	80x80	**99.63**	**99.63**	**99.63**	**99.71**	**99.63**
Ridge Regression	32x32	94.51	94.73	94.73	94.29	94.73
	48x48	97.44	98.1	98.1	97.73	97.8
	64x64	98.75	98.68	98.97	98.68	98.97
	80x80	99.19	99.12	99.12	99.19	99.12

- Ensemble selection is not necessary for the covariance-based least square estimate.
- At small window sizes, our proposed ridge estimation is performed very well. In fact, it is also robust to the variation of the numbers of coefficients per description (see Figure 5.1).