CHAPTER I

INTRODUCTION

Signal understanding abounds with different models. For example, speech is highly redundant – it is intelligible despite large distortions, and there are multiple cues for each phoneme [1,2]. Redundancy leaves room for variability of speech, in which speakers can use different feature subsets of cues. Different models in signal understanding can be different feature processing, statistical models, or search techniques. The essential idea of redundancy exploitation in signal is to estimate signal parameters in many different ways, leading to different kind of errors.

In adaptive filtering, multiple models for signal processing is proposed for *universal prediction* [3]. In this technique, the performance of a single specific model order can be improved by using a weighted combination over all possible predictors. This favorite ideas are also further applied to other fields in electrical and computer engineering communities, e.g., space-time multiantenna systems, multi-carrier code division multiple access, affine projection algorithm for least square estimation, temporal and spatial averaging for direction of arrivals, decorrelated predictor for lossless image compression, and multiple classifier systems to name a few.

Recently, a unifying conceptual framework for a variety of multiple classification algorithms, and a mathematical model of reliable transmission of information through a noisy channel has been established. This has lead to new interpretations of multiple classifier systems (MCS) which are *Error Correcting Output Code* (ECOC) [4] and its variants [5–8]. The basic idea in ECOC is to generate multiple independent decisions such that each decision (concept description) independently describes the data, and when a predefined set of descriptions is presented, they can be combined to enhance the classification.

As discussed in general classification problems [4, 8], text classification [9, 10], and automatic speech recognition [11], classification task can be viewed as a type of communication problem, where the correct category (class information) is being encoded and transmitted over a medium or channel. The channel consists of the input features, the training examples, and the learning algorithm. Because of the errors introduced by the finite training sample, poor choice of input features, and limitations or invalid assumptions made in the learning process, the class information is distorted. Regarding to the source–channel model for Machine Learning, the encoder and channel are purely conceptual, while the final classification becomes a decoding problem on the received codeword. The motivation behind the ECOC has been from mapping the output string to the nearest codeword. Most of the prior works on ECOC have been restricted to *forward error correcting* (FEC) based algorithms, which use different kinds of codes including random codes [9], repetition and

algebraic codes [4], domain and data-specific codes [5, 6, 10], and low density parity check codes (LDPC) with iterative decoding [8].

In the next section, we review several existing subspace methods of pattern recognition and relate them to multiple classifier algorithms. While much research has been done in the ECOC methods perform the decomposition a priori, the possibility of decomposing problems in feature space for improving generalization performance and interpretability has not been seriously explored. Motivated by this observation, we first propose the signal-domain channel coding approach for pattern recognition. The basic idea is to consider decomposing the feature space into a number of informative and overcomplete feature subsets, also called *descriptions*. In fact, this approach is based on the overcomplete subspace expansion, called *overcomplete wavelet representation*, widely known in signal processing community. We also propose a framework for cascading the signal-domain channel coding approach with state-of-the-art multiple classifier systems in order to achieve higher accuracy than a single multiple classifier system. We name this framework, the *generalized code concatenation* framework. Moreover, we explore several combining methods that can be used to promote a framework of learning by diversity models.

These approaches are applied to an public released MSTAR automatic target recognition and UCI repository problems. The major advantages of using these approaches are (1) improvement in generalization performance over state-of-the-art multiple classifier systems and (2) improvement in computational complexity.

1.1 Subspace Methods of Pattern Recognition

It is believed that the successful solution in pattern classification depends upon the interaction between four spaces: (1) the *input space* in which the data are available; (2) the *feature space* or the reduced input space obtained from applying any specific information preserved transform to the training data; (3) the *output space* or the exhaustive set of all the classes to which any input pattern might belong; and (4) the *hypothesis space* or a space of models in which a classifier is sought according to the training data. Most real World classification problems are characterized by large input spaces and moderately large output spaces, leading to very complex pattern classification spaces – i.e., the domain of the input space can be very complex, not only in terms of its dimensionality, but also in terms of finding manifolds of distributions of different classes within the input space when the amount of training data available is too small compared to the size of the hypothesis space; the dimensionality of the input is too large, leading to the *curve of dimensionality* problem; the learning output classes are too complex; and the true classification (or the best hypothesis) function cannot be representional due to the problems of finite available training sample, wide variety of classifier families, and available architecture choices within each family. Evidently, it is not easy to quantify the appropriateness of a hypothesis space for the given problem.

When all the spaces are too complex, it would be useful to think in some kind of space decomposition (or subspace expansion) so that each space can be represented and reconstructed from a fixed number of its simple base subspaces (most of the time, one prefers a *frame* of learning hypothesis). Learning by decomposing classification spaces can also be viewed as *a learning by diversity models*. For example, an output space can be represented by some kind of an expansion

$$h(x) = span\left(h_0(x), h_1(x), \ldots, h_k(x)\right), \tag{1.1}$$

where $h_i(x)$ denotes the i^{th} learning classifier, and x the training sample.

This subspace method for output representation can be easily explained in case of implementing ECOC method [4], where an M-valued target output function is decomposed into $k > log_2^M$ functions. This way, the components in the *span* expansion is implicitly defined by the selecting encoding scheme, while the $span^{-1}$ reconstruction function is defined by the minimum distance Hamming criterion. In boosting [12], the create of each input subset is dependent on previous classification results, and the estimated probability distribution of input space is dynamically adapted to class samples on which previous classifiers are incorrect. Multiple feature subsets method [13] can also be viewed as the subspace methods of pattern recognition. In this method, a frame of feature subsets is created for an ensemble of k-Nearest Neighbor classifiers. Subspace methods can also be easily applied to hypothesis space, when we use a wide variety of classifier families, or different architecture choices within each family.

Recently, multiple classifier systems allow us to achieve higher accuracy, which is not often achievable with single models. Evidently, most of the classification methods mentioned above are the subspace methods of pattern recognition. They are also in the family of multiple classifier systems as well. It should be noted that the necessary of multiple classifier systems is to overcome three key shortcomings of standard learning algorithms, i.e., the statistical, computational, and representational issues [14].

1.2 Motivations

The use of multiple classifier systems is motivated by their achievement in higher accuracy, which is not often achievable with single models. This achievement is come from the existence of diversity occurred in multiple predictions. The diversity of predictions is well-known for all stages in learning e.g., feature extraction, and classification.

There are at least two frameworks that can further improve generalization performance of multiple classifier systems, and are left unexplored. The first framework involves with an algorithm that can extract informative features, and at the same time, expand the feature space so that the complex true classification function can be represented and reconstructed from a fixed number of its simple base classification subspaces. Moreover, we are interested in optimal combining all the multiple classifiers obtained from the above algorithm. The

second framework is relating to the interesting question regarding to the integration of two or more multiple classifier systems in a more systematic manner, e.g., code concatenation. This dissertation attempts to provide such frameworks.

1.3 Overview of the Dissertation

We first review the basic principles of multiple classifier systems, i.e., construction approaches for multiple classifier systems (MCS) and several popular MCS, and define two important formal definitions of MCS precisely with the purpose to tie the links with signal-domain channel coding in Chapter 3, generalized code concatenation in Chapter 4, and prediction optimization method in Chapter 5.

The original contribution of this dissertation starts from Chapter 3, where we consider the application of signal-domain channel coding to multiple classifier systems. We consider an efficient feature extraction algorithm that particularly selects a basis suitable for classification from a library of orthonormal bases. There is a key observation that more resistance to overtraining is obtained when we use classifier with the basis. However, the Coiflet bases seem to be less resistant to overtraining than the bases obtained from other wavelet filters, as they are adapted too well to training data [15]. Based on this observation, we derive an algorithm that can produce multiple feature subsets (descriptions) in order to reduce the overfitting and increase the efficiency of using only one single feature basis. In fact, the technique we use here is inspired from the joint source-channel coding techniques, called *multiple description coding models*.

In Chapter 4, we derive a series of algorithms for encoding concatenated output codes based on either classical or generalized approaches in coding theory. Using two public data sets, we demonstrate the superiority of our proposed methods over the method based on a single multiple classifier system.

Chapter 5 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) always occurred in combining multiple classifiers. 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. We consider the modified ridge estimator that works effectively both in terms of computational complexity and robustness with regarding to the amount of variation caused from using different number of features.

In Chapter 6, a new method of coverage construction of multiple classifier systems is developed and applied to the public Yale face database. Several frameworks are considered and discussed for their equivalencies with an ensemble of transform networks derived using local discriminant basis algorithm. In addition to these discussions, a proof is provided that the linear combination of individual network weights of an ensemble of transform networks is a more generalized representation for multiple classifier systems than other simple methods,

e.g., constant or weighted sample mean of the weights.

Note that from this point onward, the meaning of a multiple classifier system, a classifier ensemble, a pool of classifier, an ensemble of classifiers, and a collection of classifier are the same, and these terms can be used interchangeably whenever it is appropriated. Finally, we conclude in Chapter 7 with some discussion on some further developments.

We would like to note that the following chapters are the detailed and expanded version of our published materials. Chapter 3 is based on [16, 17]. Chapter 4 is based on [18, 19]. Chapter 6 is based on [20, 21].