PSAR: PREDICTIVE SPACE AGGREGATED REGRESSION
AND ITS APPLICATION IN VALVULAR HEART DISEASE CLASSIFICATION

Ting Chen†, Ritwik Kumar†, Guillaume Troianowski⋆, Tanveer Syeda-Mahmood†, David Beymer†, Karen Brannon†

*Stanford University, Stanford, CA, USA, 94305
†IBM Almaden Research Center, San Jose, CA, USA, 95120

ABSTRACT
This paper presents a predictive space aggregated regression based boosting algorithm, and its application in classifying the Continuous Wave (CW) Flow Doppler image data set with the diseases of stenosis and regurgitation in mitral and aortic valves. The proposed algorithm involves finding a way to simultaneously combine all the weak learners based on a well-justified assumption as in the previous work [1] that not only the weak learners but each training sample should have different contributions toward learning the final strong hypothesis. However, the proposed algorithm greatly improves on the previous method by (1) dramatically reducing the number of combination weights, leading to a more stable numerical solution, (2) having regularization in both data and predictive spaces to reduce the generalization error of the model, and (3) using the sparse weight selection scheme in the testing to further avoid overfitting. A sparse subset of the training data is chosen to best approximate the test sample, and the final hypothesis is constructed based only on the chosen training samples and associated weak learner weights. Finally, we empirically show that the proposed technique not only successfully solves the overfitting problem but also significantly increases the performance of the weak classifiers via a set of comparison experiments on the CW Flow Doppler image data set consisting of 4 types of valvular diseases at different severity levels.

Index Terms— Weak classifier, Aggregated Regression, Predictive Space, CW Doppler, Valvular heart disease

1. INTRODUCTION

Echocardiography has become a key tool that is routinely used in the diagnosis and evaluation of heart disease. Valvular disease is one of the most serious heart diseases, characterized by a defect in one of the four heart valves, and may occur in newborns or at any later stage of life. Therefore, it is of great clinical importance to develop a technique for automatic valvular heart disease diagnosis based on the echocardiogram. As one of the most popular heart diagnostic tests, Continuous Wave (CW) Doppler ultrasound is widely used for the accurate assessment of blood velocity passing through the valves of the heart. Numerous varieties of classification algorithms have been proposed; however, in terms of CW Doppler image classification it is difficult to obtain good and reliable features due to large variation in the data set and intense noise in these low resolution images, including but not limited to aliasing, inhomogeneous intensity, and outliers. Researchers are interested in weak classifier ensembling techniques, such as boosting [2], bagging [3], Kernel Plurality [8], so as to attain high classification accuracy without needing to design a large feature pool. More recently, a new classifier ensembling technique CAVIAR [1] has been proposed which not only solves the overfitting problem existing in most boosting methods but also overcomes the equal weighting limitation in bagging frameworks. This algorithm achieves significantly better performances by combining extremely “weak” weak learners, however is limited in number of weak classifiers and training samples it can handle. This stems from the fact that its computational complexity a function of the product of number of weak classifiers and the number of training samples.

In this paper, we propose a novel classifier ensembling algorithm named PSAR based on predictive space aggregated regression for binary classification. Note that this framework can be easily generalized to multi-class classification by using a label vector representation (1-of-K encoding) as in [1]. We leave this for future work but only present the binary classification result in this paper. In this model, rather than assuming that the training set is composed of i.i.d. samples from the same distribution, we admit samples that are from distinct distributions. This is a better justified assumption since in most diagnostic data sets, samples belong to patients with diverse diseases, histories, age groups etc. Due to this, we allow the weights to vary with respect to the training samples. Meanwhile, we dramatically decrease the complexity of the model by reducing the number of parameters (the number of weights) from $N \times T$ to $N + T$, where $N$ is the number of training samples and $T$ is the number of weak learners. A regularization scheme is imposed so that the weights are forced to be similar for close training patterns or weak learners that behave alike. Furthermore, a sparse weight selection
scheme[4] is imposed in the testing stage to further reduce the generalization error of the model. We demonstrate the effectiveness of the proposed algorithm by comparing its performance with other existing techniques on the challenging CW Flow Doppler image data set.

2. PREDICTIVE SPACE AGGREGATE REGRESSION

Let \( D \) be the set of training samples and \( L \) be the set of labels which contains only \{+1, -1\} for binary classification. \((d_i, l_i), i = 1, \ldots, N\) are random samples drawn from the set \( D \times L \). Assume \( C_t, t = 1, \ldots, T \) are the weak classifiers applied to the instance taken from the data set \( D \). The magnitude \(|C_t|\) denotes the confidence of the prediction and its sign distinguishes the class to which it belongs. As shown in Fig.1, let \( \alpha_i, i = 1, \ldots, N \) be the weights corresponding to the training data \( d_i, i = 1, \ldots, N \), and \( \beta_t, t = 1, \ldots, T \) be the weights that are associated with the \( T \) weak classifiers. To simplify notation, denote \( \alpha = [\alpha_1, \alpha_2, \ldots, \alpha_N]^T \) and \( \beta = [\beta_1, \beta_2, \ldots, \beta_T]^T \). \( C(d_i) \) is the set of weak classifier outputs corresponding to the \( i^{th} \) training sample and \( e(C_t) \) is the set of outputs from applying the \( t^{th} \) weak classifier to all the training data. With this framework, the weight matrix that corresponds to both the training samples and weak learners is constructed through the outer product of the two weight vectors \( \alpha \) and \( \beta \), i.e. \( \alpha \otimes \beta \). Therefore, the number of parameters in this model becomes \( N \times T \) as compared to \( N \times T \) in previous work.

The goal in this classification problem is to find a proper combination of the weak learners’ outputs \( C_t \) that best approximate the ground truth label \( l_i \) for each training data \( d_i \), that is to minimize the overall error of \( \sum_{i=1}^{N}(\alpha_i C(d_i))^T \beta - l_i)^2 \). Regularization is needed to solve this ill-posed underdetermined linear system and prevent overfitting. Intuitively speaking, if the behaviors of all the weak classifiers are similar for the training data \( d_i \) and \( d_j \) (the \( i, j^{th} \) rows in Fig.1), we assume the combination strategies for both training samples should also be similar, hence the weights \( \alpha_i \) and \( \alpha_j \) should be close to each other. We can also study the behavioral similarity of each weak classifier by investigating the individual weak learner’s output from the whole training data set. If the two classifiers \( C_p \) and \( C_q \) have similar outputs, we assign similar weights \( \beta_p \) and \( \beta_q \) to them (the \( p, q^{th} \) columns in Fig.1). Let \( D(\cdot) \) be the similarity measure defined on the weak classifiers’ outputs. \( D(C(d_i), C(d_j)) \) thus represents the similarity of two data samples while \( D(e(C_p), e(C_q)) \) represents the similarity of two weak classifiers. Based on this assumption, we regularize our cost function using \( \sum_{i=1}^{N} \sum_{j=1}^{N} D(C(d_i), C(d_j))(\alpha_i - \alpha_j)^2 \) and \( \sum_{p=1}^{T} \sum_{q=1}^{T} D(e(C_p), e(C_q))(\beta_p - \beta_q)^2 \).

After the training stage, a filtering procedure is imposed during the testing stage in order to construct a data adaptive strong learner based on the given test sample, and meanwhile to reduce the generalization error of the model. As mentioned previously, we have a non i.i.d. assumption for the data set, hence only a subset of the training data may exhibit the same distribution as the test sample and only those relevant trained parameters are helpful in the testing. Recently, sparse weighted selection has been investigated for finding the most relevant training samples given a test data[4]. Following this philosophy, we formulate the testing stage of the proposed algorithm as following. Let \( A \) be a matrix with \( N \) columns where the \( i^{th} \) column is the feature vector of \( d_i \), and let \( Y = d_0 \) be the feature vector of the given test sample. We solve \( \tau^* = argmin_{\tau}||A\tau - Y||_2^2 + \nu||\tau||_1 \) to obtain the sparse combination weights \( \tau \) and construct the final hypothesis \( H \) for that certain test data \( d_0 \) using \( H(d_0) = sign(\sum_{i=1}^{N} \tau_i \alpha_i \sum_{t=1}^{T} \beta_t C_t(d_0)) \). Here \( \alpha \) and \( \beta \) are the optimal weights obtained from the training. It can be solved using the existing LASSO toolbox. Due to the \( L_1 \) norm constraint of \( \tau \), the solution is a sparse vector. The pseudo code of PSAR algorithm is listed in Algorithm 1.

Algorithm I PSAR: Predictive Space Aggregate Regression

Training stage:
1: Input \( N \) labeled training samples \( \langle (d_1, l_1), \ldots, (d_N, l_N) \rangle \), where \( l_i \in \{-1, 1\} \) and \( T \) weak learners \( C_1, \ldots, C_T \), \( C_t : D \rightarrow [-1, 1] \)
2: Initialize the weight vectors \( \alpha \) and \( \beta \) randomly
3: Minimize the following objective function:

\[
E(\alpha, \beta) = \sum_{i=1}^{N} (\alpha_i C(d_i)^T \beta - l_i)^2 + \eta_1 \sum_{i,j=1}^{N} D(C(d_i), C(d_j))(\alpha_i - \alpha_j)^2 \\
+ \eta_2 \sum_{p,q=1}^{T} D(e(C_p), e(C_q))(\beta_p - \beta_q)^2
\]

by solving for \( \alpha \) and \( \beta \) iteratively using Eqn.(2)(3) (see section 3).

Testing stage:
1: Input the test sample \( d_0 \)
2: Solve for the sparse combination weights \( \tau \)

\[
\tau^* = argmin_{\tau}||A\tau - Y||_2^2 + \nu||\tau||_1
\]

where \( A \) represents the feature vectors of the training data and \( Y = d_0 \)
3: Output the strong hypothesis

\[
H(d_0) = sign(\sum_{i=1}^{N} \tau_i \alpha_i \sum_{t=1}^{T} \beta_t C_t(d_0))
\]
3. OPTIMIZATION

In this section, we briefly discuss the optimization technique and derive the closed form solutions for solving the combination weights $\alpha$ and $\beta$. Given the cost function in Eqn.(1), we expand the right hand side of the equation and adopt the following notations: $\gamma_i = \alpha_i C(d_i); C_i = C(d_i); \epsilon_q = e(C_q); \phi_i = C(d_i)^v; \lambda_{1} = \sum_{i=1}^{N} \gamma_{i}r_i; M_{T,T} = \sum_{i=1}^{N} \gamma_{i}^2; G_{T,T} = \text{matrix has } D(d_i, d_1) + \cdots + D(d_i, d_T) \text{ in the } i^{th} \text{ diagonal}; H_{T,T} = \text{matrix has } D(d_i, d_1) + \cdots + D(d_T, d_i) \text{ in the } i^{th} \text{ diagonal}; B_{T,T} = \text{matrix has } D(d_i, d_j) \text{ in } (i,j)^{th} \text{ entry}; G', H', B' = \text{matrices with the same formats as } G, H, \text{ and } B, \text{ but defined on } D(c_i, c_j); \Phi_{N \times N} = \text{matrix has } \phi_{i,j}^2 \text{ in } i^{th} \text{ diagonal}; v_{N\times1} = [\phi_1^1, \phi_2^2, \ldots, \phi_{N}^N]'$. We developed a EM like algorithm to solve for $\alpha$ and $\beta$ iteratively until convergence.

**Algorithm 2** Solve for $\alpha$ and $\beta$

1. Fix $\beta$ and solve for $\alpha$

$$E(\alpha) = \sum_{i=1}^{N}(\alpha_i \phi_i - l_i)^2 + \eta_1 \sum_{i=1}^{N} \sum_{j=1}^{T} D(C_i, C_j)(\alpha_i - \alpha_j)^2$$

$$+ \eta_2 \sum_{p=1}^{T} \sum_{q=1}^{T} D(\epsilon_p, \epsilon_q)(\beta_p - \beta_q)^2$$

$$= \alpha^T(\Phi + \eta_1(G' + H' - 2B'))\alpha - 2\nu^{T}\alpha + \text{Const} \quad (2)$$

2. Fix $\alpha$ and solve for $\beta$

$$E(\beta) = \sum_{i=1}^{N}((\beta_i - l_i)^2 + \eta_1 \sum_{i=1}^{N} \sum_{j=1}^{T} D(C_i, C_j)(\alpha_i - \alpha_j)^2$$

$$+ \eta_2 \sum_{p=1}^{T} \sum_{q=1}^{T} D(\epsilon_p, \epsilon_q)(\beta_p - \beta_q)^2$$

$$= \beta^T(M + \eta_2(G + H - 2B))\beta - 2\lambda^T\beta + \text{Const} \quad (3)$$

Note that in both cases, the optimization problems boil down to one thing - solving the linear system $Ax = b$ with $A$ being the matrices of $\Phi + \eta_1(G' + H' - 2B')$ and $M + \eta_2(G + H - 2B)$ in the previous equations and $b$ being $v$ and $\lambda$. These can be solved with existing tools efficiently.

4. EXPERIMENTS

In this section, we empirically validate our proposed algorithm by classifying the CW Flow Doppler image data set.

**Weak Learners:** In order to demonstrate the power of our classifier ensembling algorithm, we choose the most simple weak learners by randomly selecting a dimension from the feature vector and picking a random threshold. The instances are assigned to particular classes based on their values in that chosen feature dimension by comparing to the random threshold. For example, if the feature value in that dimension is larger then the threshold, the subject is classified to class 1, otherwise, it is classified to class $-1$.

**Data Set:** In CW Doppler images, the functioning of the valves are indicated by the shapes of the Doppler signal tracings, i.e the envelopes of the velocity (see red and blue curves in Fig.2). These shape pattern of the velocity region in Doppler image has been studied in [5, 6] for the valvular disease diagnosis and it has shown promising results in decision support. Due to the normalized envelope used, the information of the severity of the diseases conveyed by the peak velocity is no longer captured. In this paper, we solve a more challenging problem of classifying the severity of the diseases by using the un-normalized envelope as well as the local intensity variation as our feature. As shown in Fig. 2, the QRS of the EKG is detected for envelope synchronization to extract one heart cycle. We obtain the velocity value by applying the optical character recognition (OCR) along the left vertical axis. For each complete heart cycle, the envelope is extracted through image processing, including smoothing, morphological operations and thresholding etc. As indicated by the black box, a local histogram is evaluated for each vertical bin contained in the envelope to capture the vertical intensity change. Both the envelope shape and the local intensity distribution information are used as our features for classification. The CW Doppler image data set contains the subjects of 51 Mitral Regurgitation(MR), 24 Aortic Stenosis(AS), 98 Mitral Stenosis(MS) and 85 Aortic Regurgitation(AR) with both EKG and envelope presented.

Each consists different degrees of severity varying from mild, moderate to severe. The typical Doppler image for each disease category is shown in Fig. 3. Note that it is trivial to classify between AR and AS (MR and MS), since the envelope shapes are very different. In the experiment, we classify between MR and AS (AR and MS), since it is a more difficult problem with unknown valve information.

**Fig. 2. Feature Extraction**

**Fig. 3. Typical images of (1) AR (2) MR (3) AS (4) MS**

**Experiment Results:** The free parameters involved in our algorithm include the regularization parameters $\eta_1, \eta_2$ in the training stage, and $\nu$ in the testing. The experiments indicate that the algorithm is not very sensitive to the choice of these parameters. In all the comparison experiments, we set $\eta_1$ and $\eta_2$ to 0.01 and $\nu = 0.5$. Euclidean distance
Table 2 shows the comparison results with the most widely used classification algorithms SVM [7] and Adaboost. For PSAR and Adaboost, we randomly take 80% as training (the rests as testing) and use 1000 weak learners. The same weak learners are used in both cases for even comparison. We repeat the experiments 100 times and report the average accuracy.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MS/AR</td>
<td>0.146</td>
<td>0.304 (51.9%)</td>
<td>0.148 (13.5%)</td>
</tr>
<tr>
<td>MR/AS</td>
<td>0.157</td>
<td>0.294 (46.6%)</td>
<td>0.358 (56.2%)</td>
</tr>
<tr>
<td>MS mild/moderate</td>
<td>0.276</td>
<td>0.368 (25.0%)</td>
<td>0.306 (9.80%)</td>
</tr>
<tr>
<td>MS moderate/severe</td>
<td>0.225</td>
<td>0.329 (31.6%)</td>
<td>0.284 (20.8%)</td>
</tr>
<tr>
<td>MR mild/moderate</td>
<td>0.334</td>
<td>0.402 (16.9%)</td>
<td>0.411 (18.7%)</td>
</tr>
<tr>
<td>MR moderate/severe</td>
<td>0.267</td>
<td>0.346 (22.8%)</td>
<td>0.364 (26.7%)</td>
</tr>
<tr>
<td>AS moderate/severe</td>
<td>0.271</td>
<td>0.201 (-34.8%)</td>
<td>0.234 (-15.8%)</td>
</tr>
<tr>
<td>AR mild/moderate</td>
<td>0.231</td>
<td>0.198 (16.7%)</td>
<td>0.338 (31.7%)</td>
</tr>
<tr>
<td>AR moderate/severe</td>
<td>0.029</td>
<td>0.042 (30.9%)</td>
<td>0.060 (51.7%)</td>
</tr>
</tbody>
</table>

The experimental results demonstrated that the PSAR algorithm significantly improves the performances of the weak classifiers and successfully solves the overfitting problem.

5. REFERENCES


