Support vector regression based autoassociative models for time series classification

S Chandrakala and C Chandra Sekhar
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai, India 600036
Email: {sckala,chandra}@cse.iitm.ac.in

Abstract—There are two paradigms for modeling varying length time series data, namely, modeling the sequence of feature vectors and modeling the sets of vectors. In this paper, we propose a regression based autoassociative model for modeling sets of vectors for time series data. We also propose a hybrid framework where a regression based autoassociative model is used for representing varying length time series data and then a discriminative model is used for classification. The proposed approach applied to speech emotion recognition task gives a better performance than the conventional methods.

I. INTRODUCTION

Modeling varying length multivariate time series data finds applications in widely different domains that involve processing data such as speech, music, video, bioinformatics, bio-medicine, and performing tasks such as speech recognition, handwritten character recognition, signature verification, speaker recognition, audio classification and speech emotion recognition. Time series data may be of discrete or continuous valued, uniformly or non-uniformly sampled, univariate or multivariate, and of fixed or varying length. Classification of time series data is a difficult task because the structure of the underlying process has to be inferred and the varying length time series have to be handled. The main issues in time series classification are related to time series representation, similarity measure and choice of classifier.

The two paradigms for modeling varying length time series data are modeling the sequences of vectors and modeling the sets of vectors. Tasks such as speech recognition need modeling both the temporal dynamics and the correlations among the features in the time series. In these tasks, production of an example belonging to a class has a fixed number of acoustic events. Hidden Markov models (HMMs) are the commonly used models for speech recognition [1]. Modeling sets of vectors includes tasks such as text independent speaker recognition, spoken language identification, audio classification, music classification and speech emotion recognition. In these tasks, duration of sequences is large and the local temporal dynamics is not critical. Production of each example of a class may have a different number of acoustic events. Gaussian mixture model(GMM) based approaches that model the sets of vectors in the time series data are used for such tasks [2]. The two main approaches to design classifiers are generative approaches [1], [2] and discriminative approaches [3], [4]. Generative approaches rely on a learned model of the joint probability distribution of the observed features and the corresponding class membership. These models are not suitable for classifying the data of confusable classes because a model is built for each class using the data belonging to that class only. Discriminative classifiers such as support vector machines (SVMs) [3] focus on modeling the decision boundaries between classes. However, these models can handle fixed length patterns only.

In this paper, we propose an autoassociative support vector regression model(AASVR) for modeling a set of vectors. We also propose a hybrid framework, that uses an autoassociative regression based method to represent a varying length time series data as a fixed length pattern and then uses a discriminative model for classification. An error vector based representation is proposed. The motivation for the hybrid framework is to make use of the advantages of similarity based paradigm for time series data representation and the discriminative approach for classification. The similarity based paradigm is shown to be effective for classification tasks [5]. In the proposed approach, each time series in the training data set is modeled by a regression based autoassociative model. The error of a time series for a given model is used as a feature. An error vector is formed by applying a time series data to the models of training data of all classes. Likewise, a test time series is also represented as an error vector. An SVM based classifier is then used for classification considering each of the error vectors as a fixed length pattern. We have used both autoassociative neural network(AANN) models and AASVR models to obtain an error vector representation of a time series data. The proposed approaches are studied for speech emotion recognition.

The rest of the paper is organized as follows: Section 2 presents a review of methods for classification of varying length time series data. Section 3 presents the regression based autoassociative models. The hybrid framework that uses an autoassociative regression based method for time series classification is explained in Section 4. Section 5 presents the studies carried out on speech emotion recognition.

II. APPROACHES TO CLASSIFICATION OF VARYING LENGTH MULTIVARIATE TIME SERIES DATA

Few approaches for modeling sets of vectors focus on finding relevant features for the task in hand. We focus on...
designing machine learning algorithms for modeling sets of vectors. The distribution of feature vectors can be modeled by parametric methods such as GMMs or nonparametric methods such as nearest neighbour approach, vector quantization based approach and neural network models. Nearest neighbour approach is employed for speaker identification task in [6]. In this work, feature vectors of registered speakers are stored as reference vectors. During testing, feature vectors of the test utterance are compared with each of the registered speaker’s feature vectors. The speaker of the test utterance is the speaker of the feature set which gives lowest distance. Vector quantization (VQ) based approach is similar to nearest neighbour modeling, except that the distance is measured to the nearest centroid which represents a cluster of feature vectors. VQ based approaches are studied in [7]. Better approach is to model the feature vectors by a set of mean and covariance parameters. This technique is employed in models such as GMM and is commonly used in modeling sets of vectors [2], [8]. GMMs make assumption about the shape of the distribution. Number of components of GMM are fixed apriori and the choice of optimal number of components is critical. Distribution of feature vectors in the feature space may not be described accurately using GMMs with its first and second order statistics and component weights. Artificial Neural Network(ANN) models do not make any assumption on the shape of the distribution. ANN models with different topologies perform different pattern recognition tasks [9], [10]. In this work, we propose regression based autoassociative models for time series classification.

III. REGRESSION BASED AUTOASSOCIATIVE MODELS

Regression is a function approximation task used to model the relation between one or more input variables and the output variables. The relation may be assumed as a linear or nonlinear function. Regression based autoassociative models (AAM) perform linear or nonlinear, identity mapping of vectors in the feature space.

A. Autoassociative neural network models

A multilayer feedforward neural network consists of interconnected processing units, where each unit represents the model of an artificial neuron, and the interconnection between two units has a weight associated with it. AANN models are multilayer feedforward neural network models that perform identity mapping. An AANN consists of an input layer, output layer and one or more hidden layers. The number of units in the input and output layers is equal to the dimension of the feature vector. The second and fourth layers of the network have more units than the input layer. The third layer is the compression layer that has fewer units than the input layer. The activation function at third layer may be linear or nonlinear, but the activation functions at the second and fourth layers are essentially nonlinear. A five layer AANN model is shown in Figure 1. Function of the five layer AANN model can be split as mapping (layers 1, 2 and 3) and demapping (layers 3, 4 and 5) networks. The mapping network projects the feature vectors in the input space \( \mathbb{R}^d \) onto an arbitrary nonlinear subspace \( \mathbb{R}^p \) formed at the compression layer. The output of the compression layer gives the reduced dimensional representation. The projection of vectors in the nonlinear subspace back onto the input space \( \mathbb{R}^d \) is performed by the demapping function. The mapping and demapping functions are nonlinear functions that play a significant role. Given a set of feature vectors of a class, the AANN model is trained using the backpropagation learning algorithm. The learning algorithm adjusts the weights of the network for each feature vector to minimize the mean squared error. It is shown in [11] that there is a relation between the distribution of the given data and the training error surface captured by the network in the input space. It is also shown that the weights of the five layer AANN model capture the distribution of the given data using a probability surface derived from the training error surface. The issues related to the architecture of AANN models are the selection of number of hidden layers and the number of units in hidden layers. Number of hidden layers and processing units in hidden layers are selected empirically. The issues related to training an AANN model with backpropagation learning are the local minima problem, suitable values for the learning rate, momentum factor and the number of iterations or the threshold as the stopping criteria. All these parameters are empirically chosen during the training. The AANN models were designed mainly for nonlinear dimension reduction. As an alternative of regression based autoassociative model, we propose an autoassociative support vector regression model for time series classification.

B. Autoassociative support vector regression (AASVR) models

The motivation for this model is to incorporate the advantages of kernel methods of regression in the autoassociative model framework. In a regression model the dependence of a scalar \( d \) on a vector \( x \) is described by \( d = f(x) + \gamma \). For a given set of training data \( \{(x_i, d_i)\}_{i=1}^{N} \), where \( x_i \) is a sample of input vector \( x \) and \( d_i \) is the corresponding value of the desired output \( d \), the problem is to provide an estimate of the dependence of \( d \) on \( x \). In support vector regression [12], nonlinear regression is considered as a linear regression by transforming data in the input space onto the high dimensional feature space. Support vector regression method approximates a function of the form

\[
f(x, w) = w^T \Phi(x) + \omega_0
\]

where \( \Phi(x) \) is the high-dimensional feature vector. For support vector regression, the error function is based on the \( \epsilon- \)
The ε-insensitive loss function defines a tube within which the error is assumed to be 0, and outside of which the error is assumed to be the difference between the absolute linear error and the radius of the tube. The constrained optimization problem is formulated by introducing two sets of nonnegative slack variables \( \{\xi_i\}_{i=1}^N \) and \( \{\xi'_i\}_{i=1}^N \) that are defined as follows:

\[
d_i - w^T\phi(x_i) \leq \epsilon + \xi_i \quad \text{for} \quad i = 1, 2, ..., N
\]

\[
w^T\phi(x_i) - d_i \leq \epsilon + \xi'_i \quad \text{for} \quad i = 1, 2, ..., N
\]

\[
\xi_i, \quad \xi'_i \geq 0 \quad \text{for} \quad i = 1, 2, ..., N
\]

where \( N \) is the number of training examples. This constrained optimization problem may therefore be viewed as equivalent to that of minimizing the cost function:

\[
\Psi(w, \xi, \xi') = \frac{1}{2} w^T w + C \sum_{i=1}^N (\xi_i + \xi'_i)
\]

subject to the constraints given above. Here, \( C \) is a user specified parameter that assigns a penalty to the error. Solving the corresponding Lagrangian function yields the dual formulation:

\[
L_d(\lambda_i, \lambda'_i) = \sum_{i=1}^N d_i (\lambda_i - \lambda'_i) - \epsilon \sum_{i=1}^N (\lambda_i + \lambda'_i) - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (\lambda_i - \lambda'_i)(\lambda_j - \lambda'_j)K(x_i, x_j)
\]

where the \( \lambda_i \) and \( \lambda'_i \) are the Lagrange multipliers. Now, the problem becomes finding the Lagrange multipliers \( \{\lambda_i\}_{i=1}^N \) and \( \{\lambda'_i\}_{i=1}^N \) that maximize the above objective function subject to the constraints:

\[
\sum_{i=1}^N \lambda_i - \lambda'_i = 0, \quad 0 \leq \lambda_i, \lambda'_i \leq C
\]

With the Lagrange optimization done, the data points for which \( \lambda_i \) or \( \lambda'_i \) are not zero are the support vectors. Finally, the optimal weight vector and the bias are computed with the help of support vectors. An estimate of desired output \( d \), denoted by \( y \), for the input vector \( x \) is obtained as follows:

\[
y = \sum_{j \geq 0} (\lambda_i - \lambda'_i)K(x, x_i)
\]

where \( N_s \) is the number of support vectors. Block diagram of the AASVR model is given in Figure 2. Unlike one model built for \( p \)-dimensional data in the AANN, autoassociative support vector regression framework needs \( p \) support vector regression models to be built. Each model is used for approximation of the function that captures the relation between the input feature vector and one feature of input feature vector. The issues in this model are the choice of the kernel and the choice of two parameters \( \epsilon \) and \( C \).

**C. Time series classification using regression based autoassociative models**

Our focus is on time series classification tasks that involve modeling sets of vectors instead of sequences of vectors. In these tasks, the duration of sequences is large and the local temporal dynamics is not critical. Production of each example of a class may have a different number of acoustic events. We assume that the feature vectors in a time series are independent. Time series having similar sets of feature vectors are assigned the same class. The GMMs are commonly used for such tasks. In case of autoassociative neural network models, one model is built for each class during training using the feature vectors of all time series in a class. In case of autoassociative support vector regression models, each class model consists of \( p \) SVR models as mentioned in the previous section. The approximate output feature vector is obtained by concatenating the output of all \( p \) models. The error for an input feature vector is the difference between the input feature vector and the approximated output feature vector. Let a multivariate time series be denoted by \( X_i = (x_{i1}, x_{i2}, ..., x_{ij}, ..., x_{in}) \), where \( x_{ij} \) is a \( p \)-dimensional feature vector and \( n_i \) is the length of the time series. Let a set of time series data, \( D = \{X_1, X_2, ..., X_M\} \), be the training data set for a class. One model is trained for each class using the feature vectors of all time series of that class. For a given test time series, \( X_i \), the average error is obtained as follows:

\[
e_i = \frac{1}{n_i} \sum_{n=1}^{n_i} \frac{\| (x_{i,n} - \hat{x}_{i,n}) \|^2}{\| x_{i,n} \|^2}
\]

During testing, average error of the test time series is calculated for each class model and the decision is based on the least average error for the test time series. Since regression is done in the kernel feature space in case of autoassociative support vector regression models, it is expected to give a better performance than AANN models. However, building AASVR model for a class of time series data that contains large number of feature vectors involves computation and storage of a large size kernel gram matrix. To address this issue, we propose a hybrid framework, that uses an autoassociative regression based method to represent a varying length sequence of vectors as a fixed length pattern and then uses a discriminative model for classification.
The motivation for the hybrid framework is to make use of the advantages of similarity based paradigm used for time series data representation and the discriminative approach for classification. In the hybrid framework, an error vector based representation is used for varying length multivariate time series data and a discriminative classifier such as SVM based classifier is used for classification. In the proposed error vector based approach, an autoassociative model (AAM) is built for each time series $X_i$ in the training data set. In case of tasks that involve modeling sets of vectors, the duration of the time series is large and hence we can afford to build one model for each time series. Then the average error is computed for each of the $M$ time series in $D$, by applying each of the $M$ models.

A time series data is now represented by an $M$-dimensional error vector that consists of $M$ error values. The generation of error vector is shown in Figure 3.

![Fig. 3. Error vector generation for a time series](image)

The steps of the proposed error vector based approach are as follows:

- Build an AAM for each time series in the training data set.
- Apply each time series to all the $M$ models and form an error vector. Each error vector $E_i$ is then normalised by $e^{(-a\cdot E_i)}$ where $a$ is a constant that is chosen empirically.
- Error vector for $X_i$ is given by $E_i = [e_{i1}, e_{i2}, ... e_{iM}]$. Such error vectors are the fixed length patterns.
- Build an SVM based classifier using the AAM error vector based representation of time series data. Apply the test time series to all the $M$ models and get the error vector.
- Use the SVM based classifier to classify the time series data represented using AAM error vector.

The dimension of the fixed length error vector depends on $M$, cardinality of training data set. Methods to mitigate the problem of higher dimensionality in case of similarity based classifiers have been suggested in [5].

V. STUDIES ON SPEECH EMOTION RECOGNITION DATA

The proposed approaches to time series classification are studied for the task of speech emotion recognition. Speech emotion recognition involves modeling the sequence of subsets of feature vectors. The Berlin emotional speech database [13] is used in our studies. Five female and five male actors uttered ten sentences in German that have little emotional content textually. The speech was recorded with 16-bit precision and at a sampling rate of 22kHz. A total of 494 utterances were divided among seven emotional classes: Neutral, Anger, Fear, Joy, Sadness, Disgust and Boredom. The duration of the utterances varies from one to two seconds. 80% of the utterances were used for training and the remaining for testing. A frame size of 20ms and a shift of 10 ms are used for feature extraction. The Mel frequency cepstral coefficient (MFCC) vector representing a given frame is a 39-dimensional vector, where the first 12 components are Mel frequency components and the 13th component is log energy. Remaining 26 components are delta and acceleration coefficients that capture the dynamics. The effectiveness of short time MFCC features in speech emotion recognition is shown in [14].

A. Studies using AANN models

The architecture of the AANN model plays a major role in capturing the distribution of the given data. The hidden units in mapping and demapping layers are responsible for capturing a nonlinear subspace and the units in the compression layer determines the number of principal components captured by the network. The number of units in the input and output layers is the same as the dimension of the feature vector. In the case of speech emotion recognition task, the number of units in the compression layer is varied from 10 to 25 and the number of units in the other hidden layers is varied from 200 to 500. The parameters of the sigmoid as activation function are bias 0.5 and slope 0.02, learning rate 0.01, momentum factor 0.5 and the number of units in the compression layer is 20. With these values, the number of units in the hidden layers other than compression layer is varied from 300 to 500 and the performance obtained is given in Table I. Best performance is obtained for the network architecture of $39L500/20L500/39L$ where $L$ denotes the linear unit and $N$ denotes the nonlinear unit. The integer values indicate the number of units in that particular layer.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>$39L300/20L300/39L$</td>
<td>49.54</td>
</tr>
<tr>
<td>$39L550/20L550/39L$</td>
<td>53.73</td>
</tr>
<tr>
<td>$39L400/20L400/39L$</td>
<td>57.03</td>
</tr>
<tr>
<td>$39L450/20L450/39L$</td>
<td>58.33</td>
</tr>
<tr>
<td>$39L500/20L500/39L$</td>
<td>58.63</td>
</tr>
</tbody>
</table>

B. Studies using error vector based approach

In case of the error vector based approach using AANN models, the architecture of the network is $39L100/20L100/39L$. The SVM based classifier is used to classify the time series data represented using the AANN based error vector. In case of error vectors obtained using AASVR models, the width parameter of the Gaussian kernel, $\epsilon$ and $C$ were empirically chosen. The performance obtained using the error vector based approaches is compared with three other methods and the results are given in Table II.

The first method is the GMM based classifier with the maximum likelihood method used for parameter estimation.
The GMM based classifier gives the best performance for 25 mixtures. Choosing optimal number of mixtures is critical in the GMM based classifier. In the second method, the variational Bayesian approach is used for parameter estimation in GMM (VBGMM). The number of mixture components of seven emotion classes (Fear(F), Disgust(D), Happiness(H), Boredom(B), Neutral(N), Sadness(S) and Anger(A)) chosen by the VBGMM models are 13,12,14,16,12,10 and 14 respectively. The VBGMM based classifier performs better than the GMM based classifier. It is shown in [2] that the VBGMM performs better for less amount of data. It automatically chooses the number of mixtures and it is free from the singularity problem that arises frequently in GMMs.

In the proposed error vector based approach using AASVR models, the kernel based nonlinear regression better approximates the input vectors in the kernel feature space and the effective discriminative ability of the similarity based paradigm also helps in achieving a better performance than that of the other methods used for comparison. Confusion matrix for the proposed error vector based approach using AASVR models is given in Table III. The examples belonging to Disgust and Anger emotion classes are correctly classified. Neutral class is confused with Boredom. The examples of Boredom class are more confusible with Disgust, Happiness and Neutral classes. The emotion specific features, instead of the MFCC features, may improve the performance of the emotion recognition system.

### TABLE II

**Comparison of Classification Accuracy (in %) for Different Approaches to Speech Emotion Recognition**

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Input to the classifier</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM</td>
<td>MFCC vector</td>
<td>64.73</td>
</tr>
<tr>
<td>VBGMM</td>
<td>MFCC vector</td>
<td>67.62</td>
</tr>
<tr>
<td>AANN</td>
<td>MFCC vector</td>
<td>58.63</td>
</tr>
<tr>
<td>SVM</td>
<td>Error vector</td>
<td>67.31</td>
</tr>
<tr>
<td></td>
<td>using AANN models</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>Error vector</td>
<td>68.57</td>
</tr>
<tr>
<td></td>
<td>using AASVR models</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE III

**Confusion Matrix (shown in %) for the Proposed Error Vector Based Approach using AASVR Models**

<table>
<thead>
<tr>
<th>Correct Emotion Class</th>
<th>Classified Emotion Class</th>
<th>F</th>
<th>D</th>
<th>H</th>
<th>B</th>
<th>N</th>
<th>S</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fear</td>
<td></td>
<td>57.14</td>
<td>0</td>
<td>18.75</td>
<td>0</td>
<td>0</td>
<td>7.69</td>
<td>0</td>
</tr>
<tr>
<td>Disgust</td>
<td></td>
<td>0</td>
<td>100</td>
<td>18.75</td>
<td>28.67</td>
<td>4.76</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Happiness</td>
<td></td>
<td>28.57</td>
<td>0</td>
<td>20</td>
<td>20.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Boredom</td>
<td></td>
<td>0</td>
<td>0</td>
<td>33.33</td>
<td>42.86</td>
<td>18.75</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Neutral</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20.00</td>
<td>52.38</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sadness</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>84.62</td>
<td>0</td>
</tr>
<tr>
<td>Anger</td>
<td></td>
<td>14.29</td>
<td>0</td>
<td>12.50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>

### VI. CONCLUSION

We have proposed an autoassociative support vector regression model (AASVR) for varying length time series classification that involves modeling sets of vectors. Building an AASVR model for a class of time series data that contains large number of feature vectors involves the computation and storage of a large size kernel gram matrix. Hence, we have proposed a hybrid framework that uses an autoassociative regression based method to represent varying length time series data as a fixed length pattern and then uses a discriminative model for classification. In this framework, the error vector based approach using the AANN models and the AASVR models is proposed. The error vector based approach using AASVR models is shown to be effective on the task of speech emotion recognition compared to the methods based on AANN models, GMM classifier and variational Bayesian GMM classifier. The proposed error vector based representation converts the difficult problem of classification of varying length multivariate time series into a problem of classification of static patterns for which well known techniques are available. The proposed representation can be applied to time series classification, clustering, matching and retrieval tasks that involve modeling sets of vectors.

### REFERENCES