Classification of Multi-variate Varying Length Time Series Using Descriptive Statistical Features

S. Chandrakala and C. Chandra Sekhar

Department of Computer Science and Engineering,
Indian Institute of Technology Madras, India
{sckala,chandra}@cse.iitm.ac.in

Abstract. Classification of multi-variate time series data of varying length finds applications in various domains of science and technology. There are two paradigms for modeling multi-variate varying length time series, namely, modeling the sequences of feature vectors and modeling the sets of feature vectors in the time series. In tasks such as text independent speaker recognition, audio clip classification and speech emotion recognition, modeling temporal dynamics is not critical and there may not be any underlying constraint in the time series. Gaussian mixture models (GMM) are commonly used for these tasks. In this paper, we propose a method based on descriptive statistical features for multi-variate varying length time series classification. The proposed method reduces the dimensionality of representation significantly and is less sensitive to missing samples. The proposed method is applied on speech emotion recognition and audio clip classification. The performance is compared with that of the GMMs based approaches that use maximum likelihood method and variational Bayes method for parameter estimation, and two approaches that combine GMMs and SVMs, namely, score vector based approach and segment modeling based approach. The proposed method is shown to give a better performance compared to all other methods.

Keywords: Time series classification, Descriptive statistical features, Speech emotion recognition, Audio clip classification.

1 Introduction

Classification of multivariate, varying length time series data is necessary in widely varying domains that include data such as speech, music, video, bioinformatics, bio-medicine and 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 real valued, uniformly or non-uniformly sampled, univariate or multi-variate and of equal or unequal length. The main issues in time series classification methods are related to (a) time series representation, (b) similarity measure and (c) choice of classifier. Approaches to time series classification focus on finding the relevant features for time series representation, or on the similarity metric between a pair of time series, and/or on modeling the time series data.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 13–18, 2009.

There are two paradigms for modeling a varying length time series, namely, modeling it as a sequence of feature vectors and modeling it as a set of feature vectors. Tasks such as speech recognition need modeling both temporal dynamics and correlations among the features in the time series. In these kind of tasks, each 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]. In tasks such as speaker recognition, audio or music classification and speech emotion recognition [2] , the duration of sequences is large, the local temporal dynamics is not critical and there may not be any underlying constraint in the time series. Each example of a class has a different number of acoustic events. Gaussian mixture models (GMMs) are commonly used for these tasks.

Generative [1,2] and discriminative approaches [3] are two main approaches to designing classifiers. Generative approaches focus on estimating the density of the data. These models are not suitable for classifying the data of confusable classes since 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 and is shown to be effective for static data classification of confusable classes. However, these models require the data to be represented as a fixed dimensional feature vector.

The motivation for the proposed work is to make use of the advantage of discriminative classifiers such as SVM for varying length time series classification tasks that involve modeling a time series as a set of vectors. In this work, we propose a method based on descriptive statistical features for multi-variate, varying length time series classification. First, local domain-specific features are extracted from each window or short time frame of a time series signal. The sequence of feature vectors is then considered as a combination of several univariate time series. A set of descriptive statistical features such as mean, variance, skewness and kurtosis are extracted from each univariate time series that forms a fixed dimensional representation for the varying length time series. The proposed representation converts the difficult problem of classification of multivariate, varying length time series into a problem of classification of static points. These static points are then classified using the SVMs. Some time series classification algorithms may fail for time series with missing samples. The proposed method reduces the dimensionality of the time series significantly and is less sensitive to the missing samples.

The rest of the paper is organized as follows: Section 2 presents a review of methods for varying length time series classification. The proposed method is presented in Section 3. Section 4 presents the results of the studies carried out on audio clip classification and speech emotion recognition.

2 Approaches to Classification of Multivariate, Varying Length Time Series Data

The two paradigms for modeling the varying length time series are modeling the sequences of vectors and modeling the sets of vectors. Approaches to the sequence

modeling of multivariate, varying length time series data can be grouped into two categories depending on the method of handling the varying length patterns. Hidden Markov model based classifiers that can handle varying length time series without changing the length or structure of the time series, form the first category of approaches [1,2]. In the second category of approaches, each time series is represented as a fixed dimensional feature vector by converting a varying length sequence of feature vectors to a fixed length pattern (a static point) so that discriminative classifiers such as SVMs can be used for classification [4].

We proposed a hybrid framework [4] for modeling sets of vectors that first uses a generative model based method to represent a varying length sequence of feature vectors as a fixed length pattern and then uses a discriminative model for classification. We proposed two approaches namely, score vector based approach and segment modeling based approach under the hybrid framework using GMM and SVM. In the score vector based approach, each time series in the training data set is modeled by a GMM. The log-likelihood of a time series for a given GMM model is used as a similarity score that forms a feature for that time series. The similarity based paradigm recently introduced for classification tasks is shown to be effective. A score vector is formed by applying the time series to all the models. Likewise, a test time series is also represented as a score vector. An SVM based classifier is then used for time series classification considering each of the score vectors as a fixed length pattern.

In score vector based representation, temporal dynamics in the time series is not modeled and the dimension of score vector depends on cardinality of training data set. In tasks such as speech emotion recognition, though local temporal dynamics is not critical, some kind of sequence information is present at gross level in the time series. To address these issues, we proposed a segment modeling based approach. In this approach, temporal ordering of segments in a time series is maintained to capture the sequence information at gross level. The parameters of a statistical model of a segment are used as features for that segment. A time series is split into a fixed number of segments. Each segment is modeled by a multivariate Gaussian model or a GMM. The model parameters of segments are concatenated in the order of the segments to form a fixed length pattern.

3 Descriptive Statistical Features Based Approach to Time Series Classification

Let a multivariate time series be denoted by $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n, ..., \mathbf{x}_N\}$, where $\mathbf{x}_j$ is a d-dimensional feature vector and N is the length of the time series. The assumption here is that the time series classification task involves modeling sets of vectors. In this proposed method, we consider a multivariate time series, $\mathbf{X}$ as a combination of d univariate time series. The i^{th} univariate time series of $\mathbf{X}$ is denoted by $\mathbf{x}_i = \{x_{1i}, x_{2i}, ..., x_{ni}, ..., x_{Ni}\}$ where x_{ni} is the i_{th} element of feature vector $\mathbf{x}_n$. We use a set of descriptive statistical features such as mean, variance, skewness and kurtosis for each of the univariate time series to describe its nature. Variance captures the degree of the deviation from the mean. Variance of a real

valued univariate time series data is the second central moment. The moments about its mean are called central moments. Normalised third central moment is called skewness. It is a measure of asymmetry of a distribution. Skewness coefficient for a univariate time series Y is

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}\frac{(y_i - \mu)^3}{\sigma^3} \tag{1}$$

where μ is the mean, σ is the standard deviation and N is the number of samples. The left skewed distribution denotes the negative skewness and the right skewed distribution denotes positive skewness. Normalised fourth central moment is called kurtosis. It is a measure of peakness of a distribution. Higher kurtosis means more of the variance is due to infrequent extreme deviations. The kurtosis for a univariate time series Y is

$$\mathbf{K} = \frac{1}{N}\sum_{i=1}^{N}\frac{(y_i - \mu)^4}{\sigma^4} \tag{2}$$

A set of descriptive features extracted from each univariate time series are then concatenated to form a fixed dimensional representation for the varying length time series. The block diagram of the proposed approach is given in Figure 1. The proposed method reduces the dimensionality of the time series significantly and is less sensitive to missing samples.

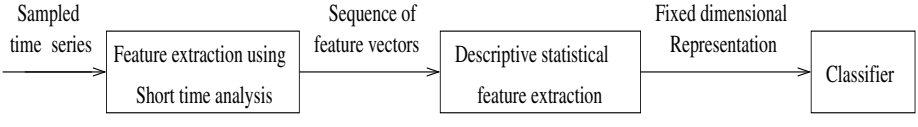

Fig. 1. Block diagram of descriptive statistical features based approach

4 Studies on Speech Emotion Recognition and Audio Clip Classification

Speech emotion recognition and audio clip classification tasks involve modeling the sets of feature vectors. Berlin emotional speech database [5] is used in our studies. 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 of the audio signal. The effectiveness of short time MFCC features in speech emotion recognition is shown in [6]. We also study classification of

audio clips of TV programs in English and Hindi belonging to the following five categories: Advertisement(Advt), Cartoon, Cricket, Football and News. Audio data is sampled at 22050 Hz. Duration of the audio clips varies from 15 to 20 seconds. A total of 987 clips were divided among five categories. 39-dimensional MFCC feature vector represents each frame of a clip. Four descriptive statistical features, namely, mean, variance, skewness and kurtosis are extracted from each of the 39 univariate time series to form a 156 dimensional vector for a time series. SVM based classifier is then used for classification. The proposed method is compared with four methods evaluated on two data sets. Table 1 shows the performance of proposed method and methods used for comparison.

Table 1. Comparison of classification accuracy (in %) of descriptive statistical features based approach with GMM, VBGMM, score vector based approach and segment modeling based approach on Berlin Emotional database and audio clip database

Classifier	Input to the classifier	Accuracy	
		Emotion Data	Audio Data
GMM	Set of MFCC vectors	64.73	90.83
VBGMM	Set of MFCC vectors	65.71	90.83
SVM	Score vector	70.48	94.80
SVM	Segment Model parameters	72.38	86.85
SVM	Descriptive statistical features	79.05	96.33

First method is the GMM classifier with maximum likelihood method for parameter estimation. In the second method, variational Bayesian approach is employed for parameter estimation in GMM (VBGMM). Since prior distributions are assumed for mixture weights, means and precision matrices instead of using point estimates, VBGMM classifier performs slightly better than GMM classifier in case of emotion data. In case of GMM score vector based representation, effective discriminative ability of similarity space helps in achieving a much better performance than GMM and VBGMM methods. The dimension of the score vector based representation depends on the cardinality of training data set. In the segment modeling based approach that uses a single Gaussian with full covariance parameters, temporal dynamics of the segments in a time series is maintained to some extent and correlations among the features within a segment are also modeled. Hence, this method performs better than the score vector based approach in case of speech emotion recognition task which involves modeling sequence of subsets of vectors. Best performance is obtained with 5 segments for emotion data and 12 segments for audio data. Segmentation of varying length time series data is a critical step in this approach. In the descriptive statistical features based method, descriptive statistical features extracted from each univariate time series effectively describe the distribution. The proposed method outperforms all other methods used for comparison for both data sets. In addition, this method is simple with less computation than all other methods. Confusion matrices for the proposed method for both data sets are given in Table 2 and Table 3.

Table 2. Confusion Matrix for the proposed method on Berlin Emotion data

Classified Class	Correct Emotion Class						
	F	D	H	B	N	S	A
(F)ear	**100**	42.86	18.75	0	4.76	0	0
(D)isgust	0	**14.29**	0	0	0	0	0
(H)appy	0	14.29	**68.75**	0	9.52	0	3.85
(B)oredom	0	14.29	6.25	**100**	4.76	30.77	0
(N)eutral	0	14.29	0	0	**80.95**	0	7.69
(S)adness	0	0	0	0	0	**69.23**	0
(A)nger	0	0	6.25	0	9.5	0	**88.46**

Table 3. Confusion Matrix for the proposed method on audio data

Classified Class	Correct Audio Class				
	Advt	Cart	Cric	Foot	News
Advt	**89.83**	2.78	0	0	1.47
(Cart)oon	5.08	**95.83**	0	0	0
(Cric)ket	0	1.39	**98.59**	0	1.47
(Foot)ball	0	0	0	**100**	0
News	5.08	0	1.41	0	**97.06**

5 Conclusion

Classification of time series data of varying length is important in various domains. Depending upon the type and nature of time series, different methods have been used to represent the time series, compute similarity between two time series and design machine learning algorithms. In this paper, a method based on descriptive statistical features for multi-variate, varying length time series classification has been proposed. The proposed method converts a difficult problem of classification of multivariate, varying length time series into a problem of classification of static patterns. It reduces the dimensionality of the data significantly, less sensitive to missing samples and involves much less computation. The proposed approach is applied on speech emotion recognition and audio clip classification. It outperforms GMM, VBGMM, score vector and segment modeling based approaches. The proposed method for representation of time series data can be applied to any time series classification, clustering, matching and retrieval tasks that involve modeling sets of vectors.

References

1. Rabiner, L., Huang, B.-H.: Fundamentals of speech recognition. Prentice Hall, NewYork (1993)
2. Mishra, H.K., Sekhar, C.C.: Variational Gaussian mixture models for speech emotion recognition. In: International Conference on Advances in Pattern Recognition, Kolkata, India (February 2009)
3. Vapnik, V.: Statistical learning Theory. Wiley-Interscience, New York (1998)
4. Chandrakala, S., Sekhar, C.C.: Combination of generative models and SVM based classifier for speech emotion recognition. In: Proc. Int. Joint Conf. Neural Networks, Atlanta, Georgia (June 2009)
5. Burkhardt, F., Paeschke, A., Rolfes, M., Weiss, W.S.B.: A database of German emotional speech. In: Interspeech, Lisbon, Portugal, pp. 1517–1520 (2005)
6. Sato, N., Obuchi, Y.: Emotion recognition using Mel-frequency cepstral coefficients. Journal of Natural Language Processing 14(4), 83–96 (2007)