Journal of Engineering and Computer Innovations Vol. 2(4), pp. 59-67, April 2011
Available online at http://www.academicjournals.org/JECI
ISSN 2141-6508 ©2011 Academic Journals

Full Length Research Paper

Sample reduction using recursive and segmented data structure analysis

R. H. Laskar*, F. A. Talukdar, Biman Paul and Debmalya Chakrabarty

Department of Electronics and Communication Engineering, National Institute of Technology, Silchar, India.

Accepted March 23, 2011

Support vector machine (SVM) is one of the widely used machine learning algorithms because of its salient features such as margin maximization and kernel substitution for classification and regression of data in a high dimensional feature space. But SVMs still face difficulties in handling large datasets. This difficulty is because of solving quadratic programming problems in SVMs which is costly, especially when dealing with large sets of training data. The proposed algorithm extracts data points lying close to the cluster boundaries of large data set, which form a much reduced but critical set for classification and regression. Inspired by the difficulties associated with SVM while handling large data sets with nonlinear kernels, the presented algorithm preselects a subset of data points and solves a smaller optimization problem to obtain the support vectors. The method presented reduces the data vectors by a recursive and segmented data structure analysis on the data vectors used to train the SVM. As this method is independent of SVM and precedes the training stage of SVM, it reduces the problem suffered by most data reduction methods that choose data based on repeated training of SVMs. Experiments using line spectral frequency (LSF) data vectors for voice conversion application show that the presented algorithm is capable of reducing the number of data vectors as well as the training time of SVMs, while maintaining good accuracy in terms of objective evaluation. The subjective evaluation result of the proposed voice conversion system is compared with the state of the art method like neural networks (NNs). The results show that the proposed method may be used as an alternative to the existing method for voice conversion.

Key words: Support vector machine, clustering based support vector machine, Mahalanobis distance, ward's linkage.

INTRODUCTION

Support vector machines (SVM) (Burges, 1998a; Vapnik, 1998g) play an important role in many areas such as pattern recognition, image processing, and many classification and regression problems. This is because of its salient properties such as margin maximization and kernel substitution for classifying the data in high dimensional feature space. Neural networks (NNs) (Haykin, 2003a) and Gaussian mixture models (GMMs) (Barrobés, 2006) are being used for classification and regression for many years. The performance of NNs depends on the training data size and network structures (Ellis et al., 1999b). As the network structure increases the training time also increases. GMM uses first and second order statistics and mixture weights, and hence, may not describe the complex distribution of the dataset appropriately. The number of mixtures should be low when there is no much data available to train the system (Mesbahi et al., 2007b).

The network structure for NNs and the number of mixtures for GMMs needs to be captured empirically. Unlike the back-propagation algorithm used to train NNs, the kernel based SVM follows the structural risk minimization problem and operates only in batch mode. The SVM with radial basis function network (RBFN) kernel best fits on the data, when number of data is large. SVMs have a small number of tunable parameters as it deals with the boundary points and is capable of finding the global solution (Burges, 1998a; Vapnik, 1998g). However, with increase in the number of data point, the limitation of SVMs becomes significant in the aspect of

*Corresponding author. E-mail: rabul18@yahoo.com.

scalability. Quadratic programming (QP) algorithms (used in SVM) are too time-consuming and memory-consuming in the case of a large number of data points.

The time-complexity and memory insufficiency problems associated with training the SVMs with large training dataset called for the need of reduced support vector machine that uses a subset of complete dataset to reduce the time-complexity and memory insufficiency problems. Many algorithms for reduction of training dataset have been proposed from time to time with their own merits and demerits (Wang et al., 2008).

Decomposition of the QP problem into several sub-problems can be used to provide a better solution to quicken SVMs training time, so that overall SVMs training time may be reduced. The time complexity can be reduced from $O(N^3)$ to $O(N^2)$ when the number of data points N is large (Wang et al., 2008). Active learning has also been applied in SVMs to select a small number of training data from the whole dataset (Hsu et al., 2002; Thong et al., 2000c; Schohn et al., 2000b). Another method uses repetitive use of SVM to find the final SVM is the incremental learning (Mitra et al., 2000a). Some random selection methods also prevail to reduce the training dataset but the probability distribution of the whole dataset is not taken into consideration in this method (Lee et al., 2001b; Watanbe et al., 2001d).

Dividing the overall quadratic programming (QP) problem of SVM into multiple QP sub-problems may quicken to solve QP problem without any extra matrix storage (Platt, 1999d). But it does not lead to a linear programming approach to reduce the complexity of the problem. The complexity may be further reduced by using Iterative Re-weighted Least Square (IRWLS) algorithm, which approximates the QP problem to a linear programming problem. The IRWLS algorithm is more computationally efficient than QP algorithms both in time and memory requirement for SVM (Pérez-Cruz et al., 2005).

Clustering based SVM (CB-SVM) (Yu et al., 2003b) method considers the clustering information in reducing the training dataset for SVM training. A clustering approach adopted in Wang et al. (2008) based on Ward's linkage is a hierarchical clustering algorithm that gives ellipsoid clusters for the complete dataset. The reduction uses Mahalanobis distance of every point within the cluster from the center of each cluster.

The data points in the vicinity of center of each cluster are removed to get reduced training dataset. But, Mahalanobis distance uses covariance matrix of the data points within a cluster. Covariance matrix is an $N \times N$ matrix, where N is number of data points whose covariance is to be calculated.

Almost all the existing methods proposed to improve the scalability of SVMs may either need to train SVMs/ repeatedly train SVMs or select randomly, select pseudo randomly or scan the whole dataset for many times to get the reduced dataset or by chunking method (Collobert et al., 2001a). These methods do not give the optimum distribution of data set and, their efficiency is still limited by the training speed of the SVMs and the scale of the dataset. Instead of using the above methods to train the SVMs, an approach proposed in Wang et al. (2008), which is based on the structure of the dataset, and is modified by recursive and segmented use of the algorithm on the dataset so that the limitation of the approach for handling large amount of data can be overcome.

The sample reduction by data structure analysis (SR-DSA) algorithm (Wang et al., 2008) was used for classification of both synthetic and real world datasets. The dataset used in Wang et al. (2008) has 2000 samples for training and 2000 samples for testing.

The proposed sample reduction using recursive and segmented data structure analysis (SR-RSDSA) which is based on SR-DSA algorithm is applied to regression problem for mapping the vocal tract characteristics of a source speaker according to that of a target speaker in a voice conversion system.

VOICE CONVERSION SYSTEM

The phenomenon of voice conversion (VC) (Barrobés, 2006; Lee, 2007a; Mesbahi et al., 2007b; Kain et al., 1998d; Stylianou et al., 1998f) is to modify a source speaker's utterance, as if spoken by a specified target speaker. The main aim of voice conversion is to design a system that can modify the speaker specific characteristics of the source speaker keeping the information and the environmental conditions, contained in the speech signal intact. In our day-to-day life, individuality in one's voice is one of the most important aspects of human speech communication. Our main objective is to design a SVM model for voice conversion which can maintain the target speaker's identity in the synthesized speech signal. Various potential applications of voice conversion are: customization of Text-To-Speech (TTS) system, developing speaker recognition and speaker verification systems in security and forensic applications, movie dubbing, animation, karaoke etc., hence the motivation for our work.

In the area of speech signal processing, isolating the characteristics of speech and speaker from the signal is a challenging problem. As the speaker identity lies in all the acoustic cues with varying degree of importance, so it is not possible to modify all the speaker specific characteristics to design a voice conversion system. The vocal tract characteristics carry the most significant information related to the identity of a particular speaker (Kuwabura et al., 1995).

The vocal tract characteristics are represented by various acoustic features, such as formant frequencies, formant bandwidths, spectral tilt (Kuwabura et al.,1995), linear prediction coefficients (LPCs) (Abe et al., 1988a), cepstral coefficients (Stylianou et al., 1998f), line spectral

frequencies (LSFs) (Arslan, 1999a), refection coefficients (RCs) (Rao et al., 2007c) and log area ratios (LARs) (Rao et al., 2007c). For mapping the speaker-specific features between source and target speakers, various models have been explored in the literature. These models are specific to the kind of features used for mapping. For instance, GMMs (Barrobés, 2006; Stylianou et al., 1998f), vector quantization (VQ) (Abe et al., 1988a), fuzzy vector quantization (FVQ) (Rao et al., 2007c), linear multivariate regression (LMR) (Baudoin et al., 1996), dynamic frequency warping (DFW) (Baudoin et al., 1996), radial basis function networks (RBFNs) (Baudoin et al., 1996), feed forward neural network (Desai et al., 2010; Srinivas et al., 2009; Rao et al., 2007c) are widely used for mapping the vocal tract characteristics.

Feature extraction and database preparation

The basic shape of the vocal tract can be characterized by the gross envelope of linear prediction (LP) spectrum. LPC parameters are obtained using the LP analysis. LSFs are derived from LPCs, which are obtained from time aligned frames (overlapping frame of 20 ms) of source and target speakers sentences. LSFs are used to describe the vocal tract characteristics, as it possesses good interpolation property. For deriving the mapping function for voice conversion, the system has to be trained with LSFs extracted from the source and the target speaker's speech signal. For this purpose, we have taken 100 parallel sentences (for training) and 30 parallel sentences (for testing) from the Arctic database of Carnegie Mellon University (CMU ARCTIC database - 0.95-release). Two male (BDL and RMS) and two female (CLB and SLT) speakers are used for this study. The CMU ARCTIC database is recorded at 16 kHz with 16 bit resolution.

To capture the relationship between the vocal tract shapes between the source and the target speakers, it needs to associate the time aligned vocal tract acoustic features of the source and the target speakers. Dynamic time warping (DTW) algorithm is used to derive the time aligned vocal tract acoustic features. Thus the database for both the source and the target speakers are prepared which consists of time-aligned 10th dimensional LSF vectors. The database contains 56054 LSF vectors for training and 23846 LSF vectors for testing the system.

Training and testing the VC system

The training of the VC system is done through VQ, NNs, and GMMs. In this paper we have used support vector regression for training the system. For a given set of input and output vectors, the goal of regression is to fit a mapping function which approximates the relation between the data points and is used later to predict the output feature space from a given new input feature space. As the database contains 10th dimensional, 56054 LSFs vectors for training, it may not possible to train the VC system with large amount of data. The SVM algorithm (Gunn, 1998c) can handle 4000 vectors and SVMTorch (Collobert et al., 2001a) can take 20000 vectors for training. The conventional methods for training the SVMs, in particular decomposition methods like SVM-Light, LIBSVM and SVMTorch handle problems with large number of features quite efficiently, but their super-linear-behavior makes their use inefficient or even intractable on large datasets (Joachim, 1999c). The algorithms mentioned in Collobert et al. (2001a), and Gunn (1998c), SVM-Light or LIBSVM needs to be applied for each dimension of the output vectors and leads to uni-dimensional regression problem. As, both the input and output feature vectors are multi-dimensional, therefore, SVM multiregressor (MSVR) (Fernandez, 2004a) is used in this study. The M-SVR can capture a nonlinear regression model which can approximate a vector-valued nonlinear function between the input and output acoustic spaces. The M-SVR can handle 4000 vectors during training, so a new algorithm named SR-RSDSA is proposed to reduce the number of training vectors to be applied to M-SVR.

Synthesis of target speaker's speech signal

The target speaker's LSF vectors corresponding to new LSF vectors (for test sentences) of the source speaker are obtained by the mapping function captured during the training phase. These predicted LSF vectors are converted to LPCs which gives the modified vocal tract characteristics of the target speaker. The modified vocal tract characteristics are excited by the source residual signal to get the target speaker's speech signal.

SUPPORT VECTOR MACHINES

Given m training pairs $(x_1, y_1), (x_2, y_2, \ldots . (x_m, y_m)$ where, $x_i \in R^d$ is an input vector labeled by $y_i \in \{+1, -1\}$ for $i = 1, \ldots \ldots, m.$ SVMs (Burges, 1998a; Vapnik, 1998f) search for a separating hyper-plane with largest margin, which is called an optimal hyper-plane $\mathbf{w^T x} + \mathrm{b} = 0$. This hyper-plane can classify an input pattern according to the following function:

$$f(x) = sgn(\boldsymbol{w}^T \boldsymbol{x} + b) \quad (1)$$

$$sgn(k) = \begin{cases} +1 \ if \ k \geq 0 \\ -1 \ if \ k < 0 \end{cases} \quad (2)$$

In order to maximize the margin for linearly separable cases, we need to find the solution for the following quadratic problem:

$$min \frac{1}{2} \| w \|^2 \quad (3)$$

subject to:

$$y_i(w^T x_i + b) \geq 1 \quad \forall i = 1,2,\ldots\ldots,m \quad (4)$$

In fact, there are many linearly non-separable problems in the real world. In order to solve these problems related to linear SVMs, we have to modify the previous method by introducing non-negative slack variables $\xi_i \geq 0$, $i = 1,\ldots\ldots,m$. The non-zero $\xi_i > 0$, are those training patterns that do not satisfy the constraints in Equation (4). The optimal hyper-plane for this kind of problem could be found by solving the following quadratic programming problem:

$$min \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{m} \xi_i \quad (5)$$

subject to:

$$y_i(w^T x + b) \geq 1 - \xi_i, \quad \forall i = 1,2,\ldots,m \quad (6)$$

$$\xi_i \geq 0 \quad (7)$$

The problem is usually posed in its Wolfe dual form with respect to Lagrange multipliers $\alpha_i \in [0, C], i = 1,2,\ldots\ldots,m$, which can be solved by standard quadratic optimization packages. The bias b can be easily calculated from any support vector x_i satisfying $0 < \alpha_i < C$. The value of α_i should be such that the support vectors can train the system effectively. If $\alpha_i = C$, then this will lead to over fitting of the system as there will be too many support vectors to handle. So for an efficient modeling, $\alpha_i's$ should be optimal. The discriminative function is therefore given by

$$f(x) = sgn(w^T x + b) = sgn(\sum_{i=1}^{m} \alpha_i y_i x_i^T x + b) \quad (8)$$

A typical SVM regression problem is to find a non linear function that is well learned by a linear learning machine in a kernel induced feature space while maintaining all the main features that characterize the maximal margin algorithm. This non linear function will then try to predict output vectors when the SVM is subjected to new input vectors. In a typical classification or regression task, only a small number of $\alpha_i's$ greater than zero are considered. The training vectors respective to $0 < \alpha_i < C$, are called support vectors, as $f(x)$ depends on them exclusively.

For some problems, improved classification or regression can be achieved using non-linear SVMs (Burges, 1998a; Vapnik, 1998f). The basic idea of nonlinear SVMs is to map data vectors from the input space to high-dimensional feature space using a non-linear mapping Φ, and then proceed for classification or regression using linear SVMs.

However, the nonlinear mapping Φ is performed by employing kernel functions $K(x_i, x)$, which obeys Mercer's conditions (Burges, 1998b), to compute the inner products between support vectors $\Phi(x_i)$ and the data vector $\Phi(x)$ in the feature space. Typical kernel functions include the radial basis function network (RBFN) $(\exp\left(-\frac{\|x-x_i\|^2}{2\delta^2}\right))$, the polynomial learning machine $((x^T x_i + 1)^p)$ and two-layer perceptron $(\tanh(\beta_0 x^T x_i + \beta_1))$ (Haykin, 2003a). For an unknown input pattern x, we have the following discriminative function:

$$f(x) = sgn\left(\sum_{i=1}^{m} \alpha_i y_i K(x_i^T x) + b\right) \quad (9)$$

In this paper, we have used RBFN kernel to project the data into feature space as RBFN implicitly calculates the bias b.

To capture the nonlinear mapping function, M-SVR (Fernandez et al., 2004a) algorithm, an approach based on IRWLS (Pérez-Cruz, 2005) algorithm is used in this study. The M-SVR was applied for non-linear channel estimation in multi-input and multi-output (MIMO) system. It is observed that M-SVR based approach provides better results in terms of bit error rate (BER) and complexity in comparison to radial basis function network (RBFN) as well as uni-dimensional support vector regression based approach.

The M-SVR is a generalization of SVR to solve the problem of regression estimation for multiple variables. The uni-dimensional regression estimation is regarded as finding mapping between an incoming vector $x_i \in R^d$ and an observable output $y_i \in R$ from a given set of i.i.d. sample $(x_i, y_i), i = 1,2,\ldots,m$. If the observable output is a vector $y_i \in R^Q$, it needs to solve the multi-dimensional regression estimation problem (Fernandez et al., 2004a). For voice conversion system, both d and Q are 10; that motivates us to use M-SVR algorithm to capture the nonlinear mapping function between the acoustic spaces of two speakers.

SAMPLE REDUCTION

Here a discussion on finding the structural information in the given dataset is carried out followed by Mahalanobis distance calculation. Thereafter, we have discussed on some of the modifications in the algorithm (Wang et al., 2008) for data segmentation and recursive reduction. The proposed algorithm for sample reduction using recursive and segmented data structure analysis (SR-RSDSA) is presented afterwards.

Data structure analysis

For many applications, whether classification or regression, data appears in homogenous groups and this structural information can provide us the basis to select the data points that are likely to be the support vectors thereby reducing the training time of the SVMs significantly.

The structure of the data is defined as the units inside which the data points are considered to share the same dispersion (Wang et al., 2008). For the purpose of investigating the structure of given dataset, hierarchical clustering (Jain et al., 1988b; Salvador et al., 2004b) is adopted to detect the clusters in each individual class. For linear SVMs, hierarchical clustering is performed in input space, and for nonlinear SVMs, in the kernel space. Specifically speaking, data points are clustered in an agglomerative manner (Jain et al., 1988b; Salvador et al., 2004b) which is described as follows:

1. Initialize each point as a cluster and calculate the distance between every two clusters.
2. While more than one cluster remains.
a. Find the closest pair of clusters.

b. Merge the two clusters.
c. Update the distance between each pair of clusters.

Hierarchical clustering approaches (Jain et al., 1988b; Salvador et al., 2004b), (single-linkage clustering, complete-linkage clustering and Ward-linkage clustering) differ in the way of finding the closest pair of clusters. The Ward-linkage clustering gives clusters that are compact and ellipsoidal, which offers a meaningful basis for the computation of the covariance matrix. If U and V are two clusters with means $\bar{U}$ and $\bar{V}$, respectively, the Ward's linkage $W(U,V)$ between clusters U and V can be calculated as (Wang et al., 2008).

$$W(U,V) = \frac{|U|.|V|}{|U|+|V|} . \| \bar{U} - \bar{V} \|^2 \quad (10)$$

Initially, each pattern is one cluster. The Ward's linkage of two patterns x_i and x_j is (Wang et al., 2008):

$$W(x_i,x_j) = \frac{\| x_i - x_j \|^2}{2} \quad (11)$$

When two clusters A and B are being merged to a new cluster A', to be more computationally efficient, the Ward's linkage between A' and cluster C, that is, $W(A',C)$ can be conventionally derived from $W(A,C)$, $W(B,C)$, and $W(A,B)$, in the following way (Wang et al., 2008):

$$W(A,C) = \frac{(|A|+|C|)W(A,C) + (|B|+|C|)W(B,C) - |C|W(A,B)}{|A|+|B|+|C|} \quad (12)$$

In the high-dimensional implicit kernel space, the hierarchical clustering is still applicable:

(1) The Ward's linkage between $\Phi(x_i)$ and $\Phi(x_j)$, that is, the images of patterns x_i and x_j, can be calculated by (c.f. Equation 10).

$$W(\Phi(x_i),\Phi(x_j)) = \frac{1}{2}[K(x_i,x_j) + K(x_j,x_j) - 2K(x_i,x_j)] \quad (13)$$

(2) Reference (Wang et al., 2008) shows when two clusters A^{Φ} and B^{Φ} merge to a new cluster A'^{Φ}, the Ward's linkage $W(A'^{\Phi}, C^{\Phi})$ between $A'^{\Phi}$and C^{Φ}can be conveniently calculated by

$$W(A'^{\Phi},C^{\Phi}) = \frac{(|A^{\Phi}|+|C^{\Phi}|)W(A^{\Phi},C^{\Phi}) + (|B^{\Phi}|+|C^{\Phi}|)W(B^{\Phi},C^{\Phi}) - |C^{\Phi}|W(A^{\Phi},B^{\Phi})}{|A^{\Phi}|+|B^{\Phi}|+|C^{\Phi}|} \quad (14)$$

During hierarchical clustering, the Ward's linkage between clusters to be merged increases as the number of clusters decreases. A curve namely the merge distance curve is drawn to represent this process. This curve is used to find the knee point (Wang et al., 2008), that is, the point of maximum curvature to determine the number of clusters.

Mahalanobis distance

As the Mahalanobis distance is calculated by utilizing the mean and variance of data statistics, it implicitly contains the data structural information. Therefore, it is more reasonable to use the Mahalanobis distance, instead of the Euclidean distance, as the distance metric. It gives the distance of a test point within the cluster from the mean of the cluster divided by width of the ellipsoid in that particular direction. It is scale invariant which means that the result will not change if all the dimensions are scaled equally.

Let X be a $n \times m$ matrix containing m random observations $x_i, i = 1,2,\ldots\ldots,m$. Let μ be the mean of the m data points and σ be the covariance, then, the Mahalanobis distance is given by

$$d^2(x_i,X) = (x_i - \mu)^T \sigma^{-1} (x_i - \mu) \quad (15)$$

If the covariance matrix is singular, it is difficult to directly calculate the inverse of σ. Instead, we can calculate the pseudo inverse σ^+ to approximate σ^{-1} as $\sigma^+ = A^T G^{-1} A$, if the Eigen-structure of the real symmetric and positive semi-definite matrix σ is $A^T G A$. Then the Mahalanobis distance from a sample x_i to the population X is:

$$d^2(x_i,X) = (x_i - \mu)^T A^T G^{-1} A (x_i - \mu) \quad (16)$$

Data segmentation

The method proposed in Wang et al. (2008) suffered from the computational complexity and memory insufficiency, due to calculation of covariance matrix and its inverse. Thus, the covariance matrix formed should be small so as to avoid memory insufficiency, which required limited amount of dataset that machine can handle. This limitation forbids the algorithm to be applied for large datasets.

To apply this algorithm for very large dataset, a small modification can be done. The data set may be divided into many segments with each having maximum number of data points whose covariance matrix and inverse of covariance matrix can be computed by avoiding the memory insufficiency problem. Then, applying the clustering algorithm to the segmented data set followed by calculation of Mahalanobis distance for the same set may help the algorithm to work for large dataset.

Recursive reduction

The data reduction obtained using the process of clustering by Ward's linkage followed by calculating the Mahalanobis distance may not be able to reduce the data to sufficient extent for large dataset. Moreover, due to segmentation chances are there that data points which form a cluster for whole dataset may not result in cluster formation for segmented data. These data points have to be kept intact because we are not sure about its contribution to support vectors. The clustering of the data is shown in Figure 1, and the corresponding reduced data set those may be the candidates for SVMs are shown in Figure 2.

The above mentioned two limitations are overcome using many stages of reduction. This recursive reduction not only reduces the dataset for training but also considers chance of the weakly clustered points in a particular stage to form cluster in subsequent stages.

ALGORITHM

The algorithm as shown in Figure 3 is based on repetitive use of the

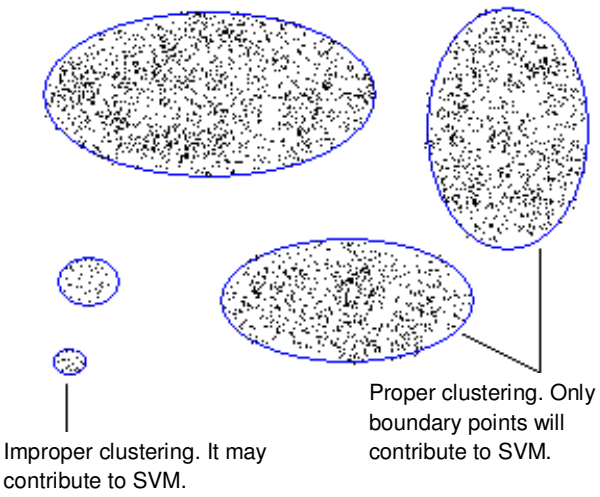

Figure 1. Proper clustering and improper clustering.

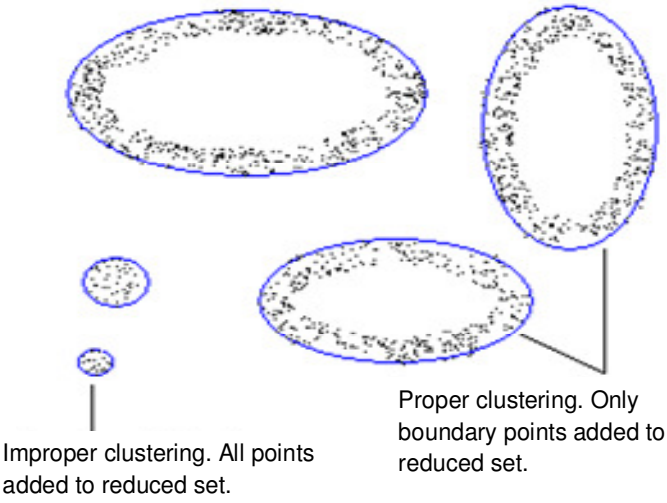

Figure 2. Reduced dataset using Mahalanobis distance.

following steps until the desired subset of dataset is obtained. The first iteration (stage or recursion) results in reduction of data to certain extent. But there may not be proper clustering of many data points because the whole data set is not considered due to memory limitation. These weakly clustered points may form proper cluster in the subsequent stages.

1.) Divide the complete training dataset into small subsets (sequentially or randomly) such that each subset contains maximum amount of data for which covariance matrix and its inverse computation does not lead to memory insufficiency problem. As the covariance matrix is an $N \times N$ data matrix for N points. This computation is a limitation to the approach if the whole set of points is fed for data reduction.
2.) If the dataset is linearly separable, the following operations can be done in the input space. If the dataset is not linearly separable, the following operations have to be done in kernel space as it becomes linearly separable in high-dimensional kernel space. Hence, for non-linearly separable dataset, it has to be projected in the kernel space.

3.) Cluster the data using hierarchal clustering technique using Ward's linkage using properly chosen cutoff distance for clustering. Hierarchal clustering using Ward's linkage takes into account the covariance of the data points and results in the formation of ellipsoidal clusters. Cutoff is properly chosen so that there is optimum amount of clustering. If the cutoff or merge distance is very less there will be many different cluster with few data points in each cluster and there would not be any assurance that the points at the vicinity of the center of the cluster shall not contribute to SVM. If cutoff is too large, the whole dataset may be clustered in single cluster. To select the optimum number of clusters the knee point of the curve between the numbers of clusters formed vs. cutoff or merge distance is chosen (Wang et al., 2008).
4.) For each cluster formed:

(a) If there is significant number of data points in the cluster, Mahalanobis distance of each point within the cluster from the mean (center of the cluster) of the data points of the cluster is calculated. η is chosen such that $1 > \eta >_0 \eta$ is multiplied with maximum Mahalanobis distance calculated for each data points within the cluster from the center of the cluster. Thereafter, the points with Mahalanobis distance less than $\eta * MAX(d_i)$ from the center are discarded.
(b) If there is very few numbers of points in the cluster, then there is no assurance that the points would contribute to SVM or not. The data is kept as it is with the hope that it might form cluster with the points that are present in the other subsets of data that were not considered at once. If left unaltered, these points may result in formation of cluster in the next stage when the above process is repeated for reduced dataset obtained for the present stage.

TESTING AND RESULTS

The M-SVR algorithm along with the SR-RSDSA is used for mapping the vocal tract characteristics of a source speaker according to that of a target speaker. LSFs derived from LPCs are used to represent the vocal tract characteristics. The training dataset contains 56054 LSF vectors. It is divided into number segments with each segment contains 2000 LSF vectors and the last segment may have less number of vectors, if number of data points is not perfectly divisible by 2000. The cutoff is chosen to be 1.1542 for male to female and 1.154 for male to male voice conversion. The value of η has been taken to be 0.8 for both the speaker combinations.

Figures 4 and 5, show the desired, predicted and source LSF vector for a particular frame of a test sentence using the above algorithm. The figures indicate that the predicted LSF vector closely follows the deisred LSF vector. The LSFs are closely related to formant frequencies and most significant information lies in the lower order formants (Barrobés, 2006; Kuwabura et al., 1995). It is observed that the proposed method for voice conversion based on M-SVR and SR-RDSA algorithm can predict the lower order LSFs more closely as compared to higher order LSFs.

The mean-square-error (MSE) is an objective measure used to evaluate the performance of the VC system. It is evaluated for the complete set of test LSF vectors.

The MSE between the desired and predicted LSF

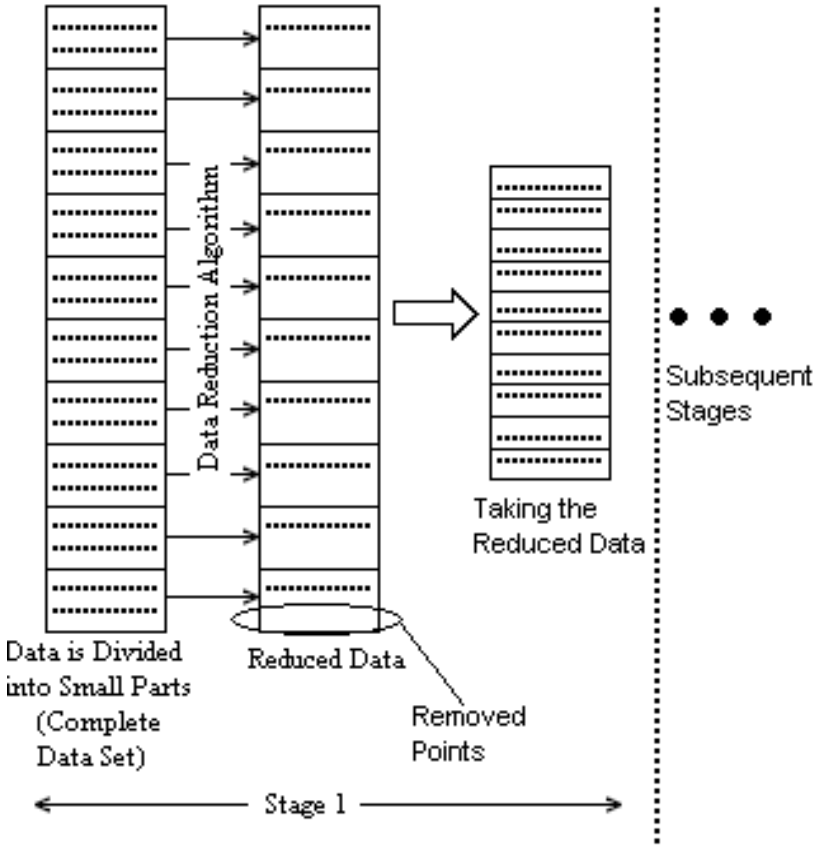

Figure 3. Proposed algorithm.

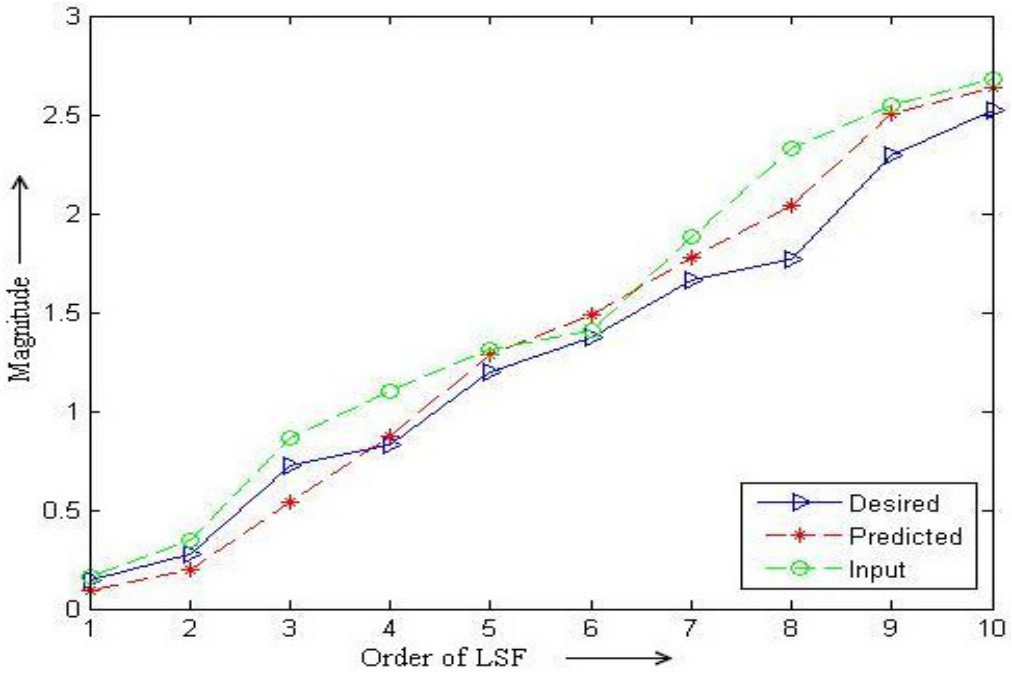

Figure 4. Desired, predicted and input LSF vector (for a particular frame of speech signal) showing the performance of the algorithm for male to female conversion. MSE is found to be 0.4 for the complete test set.

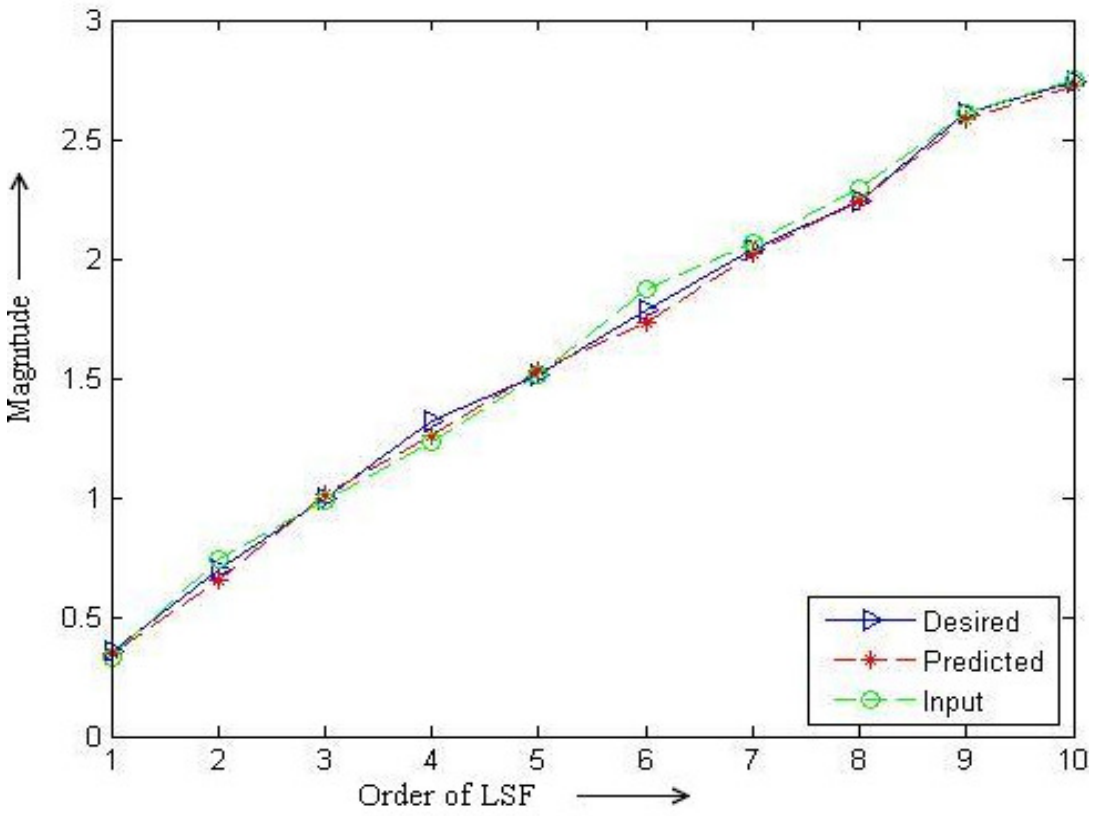

Figure 5. Desired, predicted and input LSF vector (for a particular frame of speech signal) showing the performance of the algorithm for male to male conversion. MSE is found to be 0.3 for complete test set.

Table 1. Performance evaluation interms of MOS for Voice Conversion System using M-SVR and LSF feature vectors.

Source	Target	MOS (proposed)	MOS (NN) Desai et al. (2010), Srinivas et al. (2009)
Male1	Female1	3.51	3.50
Male1	Male2	3.10	3.40

vectors is found to be 0.4 for Male 1 (BDL) to Female 1 (CLB) conversion and 0.3 for Male 1 (BDL) to Male 2 (RMS) conversion respectively.

Mean opinion score (MOS) is a subjective evaluation method used to judge the performance of a VC system. It is evaluated in a 5-point scale. The rating 5 indicates the excellent match between the original target speaker speech and the synthesized speech (that is, synthesized speech is close to the original speech of the target speaker). The rating 1 indicates very poor match between the original and synthesized utterances, and the other ratings indicate different levels of deviation between 1 and 5.

The MOS for two different speaker pair transformation is shown in Table 1. The MOS obtained using the proposed system is compared with the state art voice conversion system designed using NN and GMM (Desai et al., 2010; Srinivas et al., 2009). It may be observed from Table 1 that the proposed VC system may be used as an alternative to the existing method used for voice conversion. It is also observed that cross-gender voice conversion provides better result as compared to intra-gender voice conversion. It may due the wide differences in the vocal tarct characteristics between the two speakers belonging to different genders. The VC system designed in Desai et al. (2010), Srinivas et al. (2009) made use of CMU arctic database. However, in the design of VC system in Desai et al. (2010) Srinivas et al. (2009), mel frequency cepstral coefficients (MCEPs) is used as acoustic features to represent the vocal tract characteristics and artificial neural network (ANN) is used to train the VC system. The design of VC system using same set of acoustic features with different training algorithms may provide a better framework to evaluate

the performance of the system.

Conclusion

In this paper, data reduction for SVM using recursive and segmented data structure analysis has been proposed. The data segmentation is achieved by dividing the dataset sequentially with each segment containing maximum data set,whose covariance matrix and inverse of the covariance matrix can be calculated by avoiding memory insufficiency problem. Ward's linkage is used for hierarchal clustering so as to get compact and ellipsoidal clusters. Mahalanobis distance has been used for calculation of distance metric. Recursive reduction has been used to reduce the dataset to the desired level of reduction. M-SVR has been used for training with the reduced dataset obtained by the proposed algorithm. The predicted data points have been found to be accurate enough for the voice conversion applications. As the speaker information lies in all the acoustic features with varying degree of importance, thereby suitably transforming the other speaker specific features such as the source characteristics (shape of the glottal pulse), pitch contour, duration patterns and energy profiles may improve the performance of the system. The proposed algorithm can be applied to other data reduction application by carefully choosing the cutoff or knee point during the hierarchical clustering process.

REFERENCES

Abe M, Nakamura S, Shikano K, Kuwabura H (1988a). Voice Conversion Through Vector-Quantization. Proc. ICASSP., pp. 655-658.

Arslan LM (1999a). Speaker Transformation Algorithm using Segmental Codebooks (STASC). Speech Commun., (28): 211-226.

Barrobés HD (2006). Voice Conversion applied to Text-to-Speech Systems, PhD Thesis, Universistat Politecnica de Catalunya, Barcelona, pp. 35-94.

Baudoin G, Stylianou Y (1996). On the transformation of the speech spectrum for voice conversion. Proc. of ICSLP., pp. 1405-1408.

Burges C (1998a). A Tutorial on Support Vector Machines for Pattern Recognition. Data Min. Knowl. Discov., 2: 121-167.

Burges CJC (1998b). A Tutorial on Support Vector Machines for Pattern Recognition. Kluwer Academic Publishers, Boston, Netherlands, pp. 18-20.

Collobert R, Bengio S (2001a). SVMTorch: Support Vector Machines for Large Scale Regression Problems. J. Machine Learn., pp. 143-160.

Desai S, Black AW, Yegnarayana B, Prahallad K (2010). Spectral Mapping using Artificial Neural Network for Voice Conversion. IEEE Trans. Audio Speech Lang. Process., 18(5): 954-964.

Ellis D, Morgan N (1999b). Size matters: An Empirical Study of Neural Network Training for Large Vocabulary Continuous Speech Recognition. Proc. ICASSP, Phoenix, pp. 1-4.

Fernandez MS, Cumplido MP, García JA, Cruz FP (2004a). SVM Multi-regression for Nonlinear Channel Estimation in Multiple-Input Multiple-Output Systems. IEEE Trans. Sig. Process., 52(8): 2298-2307.

Gunn SR (1998c). Support Vector Machines for Classification and Regression. Technical Report, Department of Electronics and Computer Science, University of Southampton.

Hsu CW, Lin CJ (2002). A Comparison of Methods for Multiclass Support Vector Machines. IEEE Trans. Neural Netw., 13(2): 415-425.

Haykin S (2003a). Neural Networks - A comprehensive foundation. Prentice Hall of India, pp. 23-270.

Jain AK, Dubes R (1988b). Algorithms for Clustering Data, Prentice-Hall, New Jersey.

Joachims T (1999c). Making Large-Scale SVM Learning Practical. In B. Scholkopf, C. Burges, A. Smola (Eds.), Advances in Kernel Methods — Support Vector Learning. MIT Press, Cambridge, M.A, pp. 169-184.

Kain A, Macon M (1998d). Spectral Voice Conversion for Text-to-Speech Synthesis. Proc. ICASSP, (1): 285-288.

Kuwabura H, Sagisaka Y (1995). Acoustic Characteristics of Speaker Individuality: Control and Conversion. Speech Commun., 16: 165-173.

Lee K-S (2007a). Statistical Approach for Voice Personality Transformation. IEEE Trans. Audio Speech Lang. Process., 15: 641-651.

Lee YJ, Mangasarian OL (2001b). RSVM: Reduced Support Vector Machine. Proceedings of 1st SIAM International Conference on Data Mining, Chicago, pp. 325-361.

Mesbahi L, Barreaud V, Boeffard O (2007b). GMM-based Speech Transformation System under Data Reduction. Proceedings of 6th ISCA Speech Synthesis Workshop (SSW6), pp. 119-124.

Mitra P, Murthy CA, Pal SK (2000a). Data Condensation in Large Databases by Incremental Learning with Support Vector Machines, Proc. Int. Conf. Patt. Recognit. (ICPR2000), Barcelona, pp. 712-715.

Pérez-Cruz F, Bousoño-Calzón C, Artés-Rodríguez A (2005). Convergence of the IRWLS procedure to the support vector machine solution. Neural Comput., 17(1): 7-18.

Platt J (1999d). Fast Training of Support Vector Machines using Sequential Minimal Optimization. In: B. Scholkopf, C. Burges, A. Smola (Eds.), Advances in Kernel Methods - Support Vector Learning. MIT Press, Cambridge, MA, pp. 185-208.

Rao KS, Laskar RH, Koolagudi SG (2007c). Voice Transformation by Mapping the Features at Syllable Level. LNCS Series, Springer. 4815: 479-486.

Salvador S, Chan P (2004b). Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms. Proc. 16th IEEE Int. Conf. Tools with AI., pp. 576-584.

Schohn G, Cohn D (2000b). Less is More: Active Learning with Support Vector Machines. Proc. of 17th Int. Conf. Machine Learn. (ICML'00), pp. 839-846.

Srinivas D, Ragavendra E V, Yegnarayana B, Black A W, Prahallad K (2009). Voice Conversion using Artificial Neural Networks. Proc. ICASSP, Taiwan, pp. 3893-3896.

Stylianou Y, Cappe Y, Moulines E (1998f). Continuous Probabilistic Transform for Voice Conversion. IEEE Trans. Speech Audio Process., (6): 131-142.

Thong S, Huang DK (2000c). Support Vector Machine Active Learning with Applications to Text Classification. Proc. 17th Int. Conf. Machine Learn. (ICML'00), pp. 999-1006.

Vapnik V (1998f). Statistical Learning Theory, Wiley, New York.

Wang D, Shi L (2008). Selecting Valuable Training Samples for SVMs via Data Structure Analysis. Neurocomputing, 71: 2772-2781.

Yu H, Han J, Chang KC (2003b). Classifying Large Datasets using SVMs with Hierarchical Clusters. Proc. Int. Conf. Knowl. Discov. Data Min. (KDD'03), pp. 306-315.