A NEURAL NETWORK APPROACH FOR PREDICTING NETWORK RESOURCE REQUIREMENT IN VIDEO TRANSMISSION SYSTEMS

Hau-San Wong[†], *Min Wu*[‡], *Robert A. Joyce*[‡], *Ling Guan*[†] *and S.-Y. Kung*[‡]

[†]School of Elec. and Info. Engineering, The University of Sydney, NSW 2006, Australia.
[‡]Dept. of Electrical Engineering, Princeton University , Princeton, NJ 08544, USA.
E-mail: {hswong,ling}@ee.usyd.edu.au, {minwu,robjoyce,kung}@ee.princeton.edu

Abstract

Dynamic resource allocation is important for ensuring efficient network utilization in Internet-based multimedia content delivery system. To allow accurate network traffic prediction in the case of video delivery, relevant information based on video contents and the short term traffic pattern has to be taken into account, while the inclusion of non-relevant features will deterioriate the prediction performance due to the "curse of dimensionality" problem. In this work, we propose a neural network-based prediction system and specifically address the determination of relevant input features for the system. Experiments have shown that the current system is capable of identifying a highly relevant subset of features for traffic prediction given a large number of video content and short-term network traffic descriptors.

1. INTRODUCTION

Multimedia content delivery over the Internet requires constant monitoring and re-allocation of network resources over the bandwidth-limited channels for efficient network utilization. This is especially the case for variable bit rate (VBR) MPEG-compressed video streams, where the bandwidth requirement varies as a function of the underlying video contents. In general, video streams with low-bit rates are associated with simple scenes comprising of smooth regions and relatively little motions, while high bit rate video streams are generated from frames depicting complex scene changes. As a result, an efficient dynamic resource allocation strategy will have to take into account the current short-term traffic pattern at the bitstream level, as well as any features in the stream which implicitly describe the video contents, for effective prediction of future requirements. On the other hand, an inefficient strategy will either result in the over-allocation of bandwidths for simple scenes which leads to inefficient network utilization, or insufficient resource allocation for complex scenes which leads to buffer overflow and packet loss.

In view of the above requirements, we propose the adoption of both video content features and short-term traffic data for predicting future resource requirements. Specifically, we partition a video stream into individual camera shots [1], and the bandwidth renegotiation points are selected near the beginning of each shot where the above features are observed for a short time interval and these are then used for predicting the required resources for the entire shot. In other words, this strategy is based on the assumption that the traffic patterns are relatively constant within a single shot. The content features in the form of a set of DCT coefficient and motion vector statistics are extracted directly from the compressed domain to allow low-delay processing. For the short term traffic data, we adopt the measurements deriving from the deterministic bounding interval dependent (D-BIND) model proposed by Knightley *et.al* [2], which describes the maximum allowed bit arrival rate for time intervals of various lengths.

Due to the possible complex dependencies between the long-term traffic patterns and the set of content and short-term D-BIND features, the prediction system is implemented in the form of a multilayer perceptron which is capable of approximating complex mappings through a training process [3]. Among the complete set of features, it is also apparent that not all of them will be relevant for traffic prediction. In fact, in the case of a limited number of training samples, the inclusion of non-relevant features will lead to the "curse of dimensionality" problem where the sparseness of the samples in a high dimensional space will lead to incorrect reconstruction of the desired mapping [3]. In view of this, we propose the adoption of the sequential forward selection (SFS) approach [4] which incrementally constructs a sequence of feature subsets by successively adding relevant features to those previously selected. Due to the need for evaluating the relevancy of the subsets, which normally requires iterative identification of multiple mappings between the corresponding reduced feature spaces and the long term traffic patterns, we adopt the general regression neural network (GRNN) [5] for this purpose. Unlike alternative neural network models which require iterative learning, the parameters of the GRNN model can be directly determined in a single pass of training, which allows rapid evaluation of the individual feature subsets in terms of their relevancies.

2. CONTENT AND TRAFFIC DESCRIPTORS

The original set of features is denoted as the index set $F = \{1, \ldots, N\}$. In this work, the set F consists of 14 content descriptors and 4 short-term traffic descriptors in the form of D-BIND measurements. In other words, $N = 18$ and the features are numbered in such a way that the subset $F'_C = \{1, \ldots, 14\}$ contains the content descriptors and the subset $F'_D = \{15, \ldots, 18\}$ contains the D-BIND descriptors. For the content descriptors, we have included features which describe both the spatial and temporal characteristics of the video. Among those in the first category is the spatial complexity measure associated with the I-frames in the MPEG stream, which is defined as a specific weighted sum of the AC DCT coefficient magnitudes in each block of the frames. Features describing the temporal characteristics include the mean and variance of the magnitudes of motion vectors (MV), the covariance between the x and y-components of the vectors, mean change in MV magnitudes over 2 P-frames, and related features.

For the short term traffic descriptors, the D-BIND measurements proposed by Knightley *et.al* [2] are adopted which are defined as follows: if $A[\tau, \tau + t]$ denotes the total number of bits received during the interval $[\tau, \tau + t]$, we can define the *empirical envelope* $B^*(t)$, which is a function of the interval length t, as the least upper bound of the number of received bits over all possible values of τ, i.e.,

$$B^*(t) = \sup_{\tau} A[\tau, \tau + t] \tag{1}$$

Given a specific discretization of the allowed time interval lengths $t_l, l = 1, \ldots, L$, the empirical envelope function can be described as a vector $\mathbf{B} = [b_1, \ldots, b_l, \ldots, b_L]^T$ where $b_l = B^*(t_l)$. The D-BIND descriptor is then defined as the associated vector $\mathbf{D} = [r_1, \ldots, r_l, \ldots, r_L]^T$, where the components $r_l, l = 1, \ldots, L$ denote the bit arrival rates $r_l = b_l/t_l$ for each possible interval length t_l. Given the D-BIND vector $\mathbf{D}$, we can thus characterize the short-term network traffic by those components r_l with small indices l, and the long-term traffic pattern by those r_l associated with large indices. Our objective is to predict the long term D-BIND components from a judicious combination of the short-term D-BIND components and the previous content descriptors, such that an accurate estimation of the required bandwidth over an extended period can be achieved by briefly observing the content and traffic characteristics at the beginning of that period.

Formally, a chosen subset of the content and traffic descriptors are regarded as input features to the desired prediction system. While we can choose a set of r_l values with large l indices from the D-BIND vector and designate them as the desired output values, it was observed that most of these component values are close to the average bit rate in the interval and are thus redundant. For more efficient description, we perform principal component analysis on the set of D-BIND vectors associated with all the video shots, and then adopt the projections of each D-BIND vector on to the first two principal components as the long-term traffic descriptors.

3. FEATURE SELECTION

The importance of selecting the relevant subset from the original feature set is closely related to the "curse of dimensionality" problem in function approximation, where sample data points become increasingly sparse when the dimensionality of the function domain increases, such that the finite set of samples may not be adequate for characterizing the original mapping [3]. In addition, the computational requirement is usually greater for implementing a high-dimensional mapping. To alleviate these problems, we reduce the dimensionality of the input domain by choosing a relevant subset of features from the original set. Specifically, we propose the adoption of an efficient nonlinear one-pass selection procedure, the sequential forward selection (SFS) method, and a specialized neural network model, the general regression neural network (GRNN) [5] to achieve this purpose. Given the large number of possible feature combinations from the original set, the former represents a sub-optimal yet efficient approach of sampling from the space of possible feature subsets such that those which characterize the original mapping with reasonably good accuracy can be quickly identified. In addition, due to the need to apply iterative function approximation techniques for mapping each candidate subset to the desired function outputs for relevancy evaluation, the adoption of the GRNN model provides an alternative approximation approach which, unlike other NN models, requires only a single pass of training to allow rapid evaluation of the candidate subsets in terms of their relevancies.

3.1. Sequential Forward Selection (SFS)

For feature selection, we formally denote the original set with N features as the index set $F = \{1, \ldots, N\}$. Our purpose is to approximate the original mapping $f : D_F \subset \mathbf{R}^N \longrightarrow \mathbf{R}^S$, where $\mathbf{x}_F = [x_1, \ldots, x_N]^T \in D_F$ are the vectors in the input domain, using the alternative mapping $g : D_{F'} \subset \mathbf{R}^M \longrightarrow \mathbf{R}^S$, where $F' = \{i_1, \ldots, i_M\} \subset F$ is the relevant feature subset with $M \leq N$, and $\mathbf{x}_{F'} = [x_{i_1}, \ldots, x_{i_M}]^T \in D_{F'}$ are the vectors in the associated function domain. To achieve this, the subset F' has to be chosen in such a way that $f(\mathbf{x}_F) \approx g(\mathbf{x}_{F'})$ for every $\mathbf{x}_F \in D_F$ and $\mathbf{x}_{F'} \in D_{F'}$.

The sequential forward selection procedure [4] allows construction of a suitable feature subset starting from a single feature. Specifically, given the original feature set F, the SFS algorithm generates a sequence of subsets $F'_m, m = 0, \ldots N$ with associated cardinalities $|F'_m| = m$. To begin with, we require a measure to evaluate the relevancies of the candidate subsets F'_m. For this purpose, we are usually given a set of training data $(\mathbf{x}_{F,p}, \mathbf{y}_p), p = 1, \ldots, P$ for the desired mapping, where $\mathbf{x}_{F,p} \in D_F \subset \mathbf{R}^N$ denotes each sample vector in the input domain incorporating the full set of features. To evaluate the relevancy of a particular feature subset F'_m, we construct a mapping $g_{F'_m}$ which minimizes the following mean square error measure

$$E_{F'_m} = \frac{1}{P} \sum_{p=1}^{P} \|\mathbf{y}_p - g(\mathbf{x}_{F'_m,p})\|^2 \tag{2}$$

over the set of reduced dimension vectors $\mathbf{x}_{F'_m,p} \in D_{F'_m} \subset \mathbf{R}^m$. We can then regard the value $E_{F'_m}$ as a measure of the relevancy of F'_m.

Given the m-th feature subset $F'_m = \{i_1, \ldots, i_m\}$, we generate the $(m+1)$-th relevant subset by individually evaluating the suitability of each remaining feature in the complement set $\overline{F'}_m = F - F'_m = \{i_{m+1}, \ldots, i_N\}$. For each remaining feature $i_j, j = m+1, \ldots, N$, we form the new subset $G'_{m+1,j}$ as follows:

$$G'_{m+1,j} = F'_m \cup \{i_j\}, j = m+1, \ldots, N \tag{3}$$

The $(m+1)$-th relevant subset is then chosen from these candidate subsets using the following criterion:

$$F'_{m+1} = G'_{m+1,j^*} \text{ where } j^* = \arg\min_j E_{G'_{m+1,j}} \qquad (4)$$

In this way, a nested sequence of feature subsets $F'_1 \subset \ldots \subset F'_N$ can be constructed, and the associated performance measure values $E_{F'_1}, \ldots, E_{F'_N}$ indicate the relevancy of the corresponding subsets. As a result, we can select the subset containing the minimum number of features and with its associated measure value lower than a prescribed threshold.

3.2. General Regression Neural Network (GRNN)

The general regression neural network (GRNN) [5], which is shown in Fig 1. is a special example of a radial basis function (RBF) network [3], in which the centers and widths of the Gaussian kernels are represented as deterministic functions of the training data. In other words, no iterative training procedures are required to reconstruct a mapping. Since one of the problems with the feature selection process is the necessity to reconstruct $g_{F'_m}$, usually by an iterative function approximation process, for the evaluation of $E_{F'_m}$ in Eq (2), the adoption of GRNN will allow rapid evaluation of the relevancy of different feature subsets for our current problem.

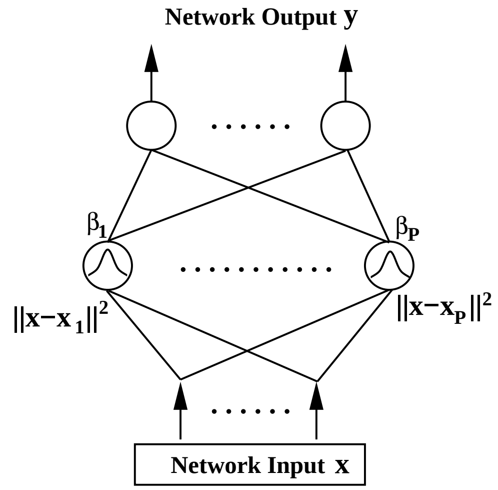

Figure 1: Architecture of GRNN

To carry out function approximation using GRNN, we are given a set of sample observations $(\mathbf{x}_p, \mathbf{y}_p), p = 1, \ldots, P$ from the original function. We then assign each input vector $\mathbf{x}_p$ as the center of a corresponding Gaussian kernel in the network. For an arbitrary input vector $\mathbf{x}$, the output of the p-th RBF unit is given by

$$\beta_p = \exp\left[-\frac{(\mathbf{x}-\mathbf{x}_p)^T(\mathbf{x}-\mathbf{x}_p)}{2\sigma^2}\right] \qquad (5)$$

where σ is a user-specified smoothing parameter. The estimated function output $\mathbf{y}$ for $\mathbf{x}$ is then given by the following convex combination:

$$\mathbf{y} = \sum_{p=1}^{P} \alpha_p \mathbf{y}_p, \quad 0 \le \alpha_p \le 1, \sum_{p=1}^{P} \alpha_p = 1 \qquad (6)$$

where the coefficients α_p are defined as $\alpha_p = \beta_p / \sum_{p=1}^{P} \beta_p$, $p = 1, \ldots, P$. Intuitively, the GRNN performs interpolation by linearly combining the given training outputs $\mathbf{y}_p$. If the current input vector $\mathbf{x}$ is close to one of the training input $\mathbf{x}_p$ in a Euclidean sense, the corresponding coefficient α_p given by Eq (6) will also become large, and the estimated output $\mathbf{y}$ will be close to the desired output $\mathbf{y}_p$ for $\mathbf{x}_p$, which is a reasonable construction. On the other hand, those sample points which are far away from $\mathbf{x}$ do not appreciably contribute to the summation due to the exponentially decaying weighting function α_p.

4. EXPERIMENTAL RESULTS

In this section, we apply the SFS technique to select feature subsets from the original set of 4 short-term D-BIND features and 14 content features, and the GRNN to evaluate their relevancies. Our experiments are performed on a set of 3 video sequences digitized from cable television at 30 frames per second. Using the automatic shot boundary detection algorithm in [1], 177 shots are identified and features associated with each shot are extracted. We plot the error values for each subset F'_m in Fig. 2, where the numbers in the horizontal axis correspond to the index of the new features selected in each SFS trial. It is seen that the error curve exhibits a distinct minimum point at the feature subset F'_6, which in our case corresponds to the index subset $\{1, 6, 15, 16, 17, 18\}$. A possible interpretation of this minimum is that, given the limited number of training samples and their increasing sparseness in high-dimensional spaces, the simple GRNN model, which does not incorporate any explicit trainable parameters, will find it increasingly difficult to characterize the mapping beyond a certain maximum number of features, and the error starts to rise beyond this point. As a result, it is natural to adopt F'_6 as a first approximation to our relevant feature subset due to the associated small error value and the limited training set size. Among the selected features, it is observed that all the four short-term D-BIND descriptors are included, which implies that the short-term traffic statistics are essential for predicting the long-term traffic patterns. It is next observed that the complexity feature associated with the I-frames (feature 1) and the mean change in magnitude of the motion vectors over 2 P-frames (feature 6) are also important for prediction.

To further verify the feature selection results, we implement a complete traffic prediction system in the form of a multilayer perceptron. The input consists of the previously selected features observed over a short period at the beginning of each video shot. The output describes the long-term traffic pattern in the form of the first two principal component projections of the complete D-BIND vector over the entire shot. The back propagation (BP) algorithm is applied to determine the weights and biases of the network. Among the 177 shots, the first 50 shots are used for training and the rest for testing. We have listed

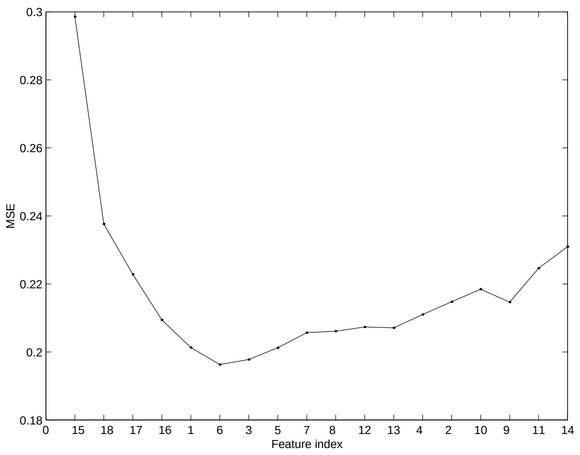

Figure 2: Error plot for SFS/GRNN feature selection

the prediction mean square error in normalized units for different number of hidden nodes in Table 1. For the purpose of comparison, we have also included the prediction results by randomly choosing 2 sets of 6 features from the original 18 (indicated by RS1 and RS2 in the table). We can observe that the 6 features selected by SFS and GRNN achieve the smallest error in each case. In addition, we also notice that increasing the number of hidden nodes from 10 to 20 does not significantly improve the prediction results, and for some particular feature combinations the prediction error even increases for a large hidden layer, indicating the possibility of overfitting.

Subset	no.of units	MSE (1st PCA)	MSE (1st+2nd PCA)
F'_6	10	0.0238	0.0277
RS 1	10	0.0559	0.0695
RS 2	10	0.0426	0.0545
D-BIND	10	0.0247	0.0281
F'_6	20	0.0232	0.0268
RS 1	20	0.0579	0.0719
RS 2	20	0.0488	0.0617
D-BIND	20	0.0244	0.0279

Table 1: Prediction results using MLP

From these results, we can conclude that the SFS/GRNN selection mechanism is capable of identifying the most important features, in the form of the short-term D-BIND statistics, for the current prediction problem. On the other hand, we can observe that the addition of content features to the D-BIND subset serve to refine the prediction result.

Since the ranking of all D-BIND features are close to the top of the feature list, it is reasonable to suggest that most of the useful information for predicting the future traffic is already embedded in these short-term statistics. To confirm this, we have also included the prediction results using the 4 short-term D-BIND features only. We can observe that the resulting errors are only slightly greater than those of the original selected subset F'_6, indicating that these short-term features are the most essential for predicting the long term network traffic.

5. CONCLUSION

We have proposed a neural network-based traffic prediction strategy for dynamic resource allocation in video transmission systems. Specifically, the problem of determining the relevant input features for the prediction system is addressed. For this purpose, we adopt the SFS method to construct a sequence of feature subsets, and the GRNN to allow the efficient evaluation of the relevancies of these subsets without requiring iterative training. The full prediction system is implemented in the form of a multilayer perceptron with the previously selected features as network inputs. From the experimental results, it was observed that the combined SFS/GRNN selection strategy is capable of identifying the short-term D-BIND statistics as the most important features for the current prediction problem. It was also seen that the addition of content features serve to refine the prediction result. For future research, we propose to investigate complementary approaches for selecting additional content features along the lines of [6] and [7] for further prediction improvement. This is in view of our observation that, during the feature selection process, only a comparatively small number of content features are currently included due to the sensitivity of the simple GRNN model to the "curse of dimensionality" problem beyond the minimum point in Fig. 2, while it may be the case that some of the remaining content features will be useful for prediction in the full system.

6. REFERENCES

[1] B. L. Yeo and B. Liu, "Rapid scene analysis on compressed video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 6, pp. 533–544, 1995.

[2] E. W. Knightly and H. Zhang, "D-BIND: An accurate traffic model for providing QoS guarantees to VBR traffic," *IEEE/ACM Trans. on Networking*, vol. 5, no. 2, pp. 219–231, 1997.

[3] S. Haykin, *Neural Networks: A Comprehensive Foundation.* NJ: Prentice Hall, 1999.

[4] J. Kittler, "Feature set search algorithms," in *Pattern Recognition and Signal Processing* (C. H. Chen, ed.), Sijthoff & Noordhoff, 1978.

[5] D. F. Specht, "A general regression neural network," *IEEE Trans. Neural Networks*, vol. 2, no. 6, pp. 568–576, 1991.

[6] P. Bocheck and S. F. Chang, "Content-based VBR traffic modeling and its application to dynamic network resource allocation," Tech. Rep. 48c-98-20, Columbia University, 1998.

[7] M. Wu, R. Joyce, and S. Y. Kung, "Dynamic resource allocation via video content and short-term traffic statistics." to be presented in IEEE Int. Conf. on Image Processing, 2000.