International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

Web Mining of High Dimensional Data Streams Using Tensor Analysis – A Predictive Study using MATLAB

Anupma Sinha
Department of CSE
Shri Venkateshwara University
sinhaanupma3@gmail.com

Dr. Anshu Srivastava
Department of CSE
Shri Venkateshwara University
anshuqrat114@gmail.com

***Abstract*--Tensor decompositions are very versatile and powerful tools, ubiquitous in data mining applications. As seen, they have been successfully integrated in a rich variety of real world applications, and due to the fact that they can express and exploit higher-order relations in the data, they tend to outperform approaches that ignore such structure.**

The work deals with the applicability and capability of Tensor Analysis techniques for time series analysis of network traffic response data prediction has been investigated. It is seen that the models are very robust, characterised by fast computation, capable of handling the noisy and approximate data that are typical of data used here for the present study. Due to the presence of non-linearity in the data, it is an efficient quantitative tool to predict defect in softwares. The studies has been carried out using MATLAB simulation environment. From the analysis of the results it is seen that the network traffic response prediction model developed using subtractive clustering technique has been able to perform well.

Keywords: Tensor, Data Mining, Network Traffic, MATLAB

1. Introduction

Tensors are multidimensional extensions of matrices. Because of their ability to express multimodal or multiaspect data, they are very powerful tools in applications that inherently create such data. For instance, in online social networks, people tend to interact with each other in a variety of ways: they message each other, they post on each other's pages, and so on. All these different means of interaction are different aspects of the same social network of people, and can be modeled as a three-mode tensor, a "data cube," of (user, user, means of interaction). Given this tensor, there exists a rich variety of tools called tensor decompositions or factorizations that are able to extract meaningful, latent structure in the data.

Tensor decompositions are very versatile and powerful tools, ubiquitous in data mining applications. As seen, they have been successfully integrated in a rich variety of real world applications, and due to the fact that they can express and exploit higher-order relations in the data, they tend to outperform approaches that ignore such structure.

Furthermore, recent advances in scaling up tensor decompositions have employed practitioners with a strong arsenal of tools that can be applied to many big multi-aspect data problems. The success that tensors have experienced in data mining during the last few years by no means indicates that all challenges and open problems have been addressed. They have been used for modelling space and time, unsupervised model selection, dealing with higher order data and connection with heterogeneous information networks. In the present work applicability and capability of Tensor Analysis techniques for time series analysis of network traffic response data prediction is going to be investigated.

2. Data Used

The web log data file available from the data repository at http://ita.ee.lbl.gov/html/contrib/EPA-HTTP.html is being used here for web usage mining. Every line in log file represents a request made by a browser of user. Fig.1 shows the web log file.

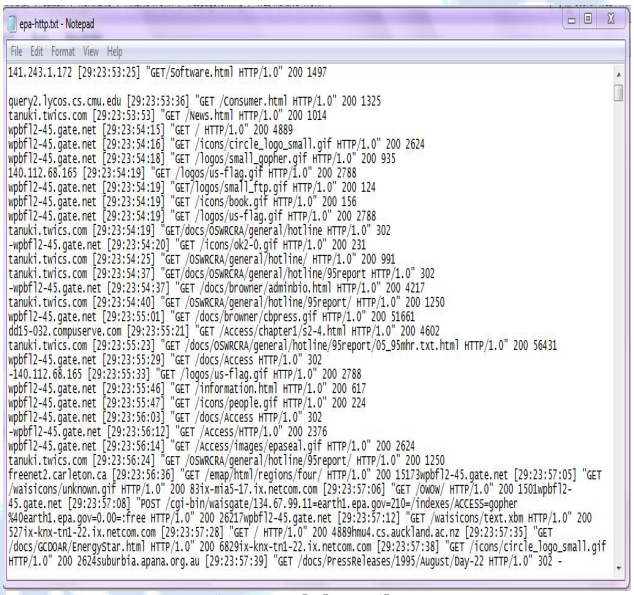

Fig.1. Web log File.

3. Traffic modeling and analysis of Web Data Using Tensor Analysis

3.1 Preprocessing of Data for ICA

Generally, ICA is performed on multidimensional data. This data may be corrupted by noise, and several original dimensions of data may contain only noise. So if ICA is performed on a high dimensional data, it may lead to poor results due to the fact that such data contain very few latent components. Hence, reduction of the dimensionality of the

data is a preprocessing technique that is carried prior to ICA. Thus, finding a principal subspace where the data exist reduces the noise. Besides, when the number of parameters is larger, as compared to the number of data points, the estimation of those parameters becomes very difficult and often leads to over-learning. Over learning in ICA typically produces estimates of the independent components that have a single spike or bump and are practically zero everywhere else [5]. This is because in the space of source signals of unit variance, nongaussianity is more or less maximized by such spike/bump signals. Apart from reducing the dimension, the observed signals are centered and decorrelated. The observed signal X is centered by subtracting its mean:

$$X \leftarrow X - E\{X\} \quad (1)$$

Second-order dependences are removed by decorrelation, which is achieved by the principal component analysis (PCA) [6,7]. The ICA problem is greatly simplified if the observed mixture vectors are first whitened. A zero-mean random vector z = (zi .....zj)T is said to be white if its elements z are uncorrelated and have unit variances E{zi zj } = $\delta_{i,j}$

In terms of Covariance matrix, the above equation can be restated as,

$$E\{zz^T\} = I \quad (2)$$

where I is the identity matrix. A synonymous term for white is sphered. If the density of the vector z is radially symmetric and suitably scaled, then it is sphered, but the converse is not always true, because whitening is essentially decorrelation followed by scaling, for which the PCA technique can be used.

3.2 Algorithms for ICA

Some of the ICA algorithms require a preprocessing of data X and some may not. Algorithms that need no preprocessing (centering and whitening) often converge better with whitened data. However, in certain cases, if it is necessary, then sphered data Z is used. There is no other mention of sphering done for cases where whitened is not required. 5.1 Non-linear Cross Correlation based Algorithm The principle of cancellation of nonlinear cross correlation is used to estimate independent components in [10,11]. Nonlinear cross correlations are of the form E{g1 (yi)g2 (yj)}, where g1 and g2 are some suitably chosen nonlinearities. If yi and yj are independent, then these cross correlations are zero for yi and yj , having symmetric densities. The objective function in such cases is formulated implicitly and the exact objective function may not even exist. Jutten and Herault, in [5], used this principle to update the non diagonal terms of the matrix W according to

$$\Delta W_{ij} \propto g1(yi)\,g2(yj) \text{ for } i \neq j \quad (3)$$

Here yi are computed at every iteration as y = (I + W)−1z and the diagonal terms Wij are set to zero. After convergence, yi give the estimates of the independent components. However, the algorithm converges only under severe restrictions [12].

4 Modelling of Network Traffic Data

Traffic modeling and analysis plays an important role in determining network performance. A model which can accurately interpret the important characteristics of traffic is required for efficient analytical study and simulation purposes. This in turn triggers better knowledge of network dynamics, thus an essential aid for network design and bandwidth wastage control. Traffic modeling originated from the study of telephone network with the Poisson assumption of the traffic arrival process. However, with the emergence of modern technology, network traffic has evolved tremendously becoming more complex and bursty than the earlier voice traffic. This resulted in several sophisticated stochastic models. But to accurately predict a network traffic, there is a need for traffic models which can capture the characteristics. Network engineering and management rely a lot on an appropriate model for traffic measurements. Traffic forecasting is one major research interest for many network engineers.

4.1 Model Inputs and Structure

The modeling approach used to develop prediction model along with details on input and output parameters is presented in this section. One of the most important steps in the development of any prediction model is the selection of appropriate input variables that will allow the soft computing technique to successfully produce the desired results. Good understanding of the system under consideration is an important prerequisite for successful application of data driven approaches. Physical understanding of the process being studied leads to better choice of the input variables. Here since predicting defect in software is a complicated problem that involves multiple interacting factors.

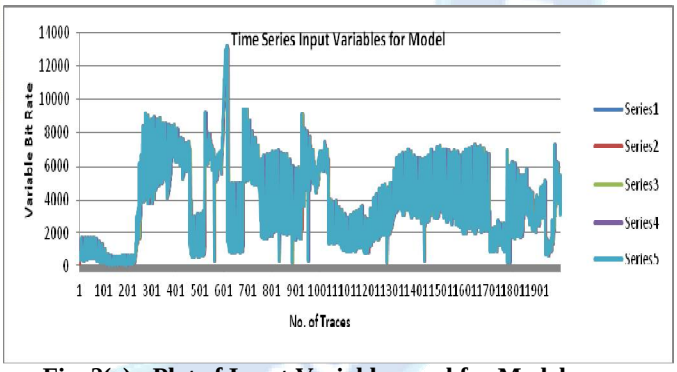

Fig. 2(a) : Plot of Input Variables used for Model Development

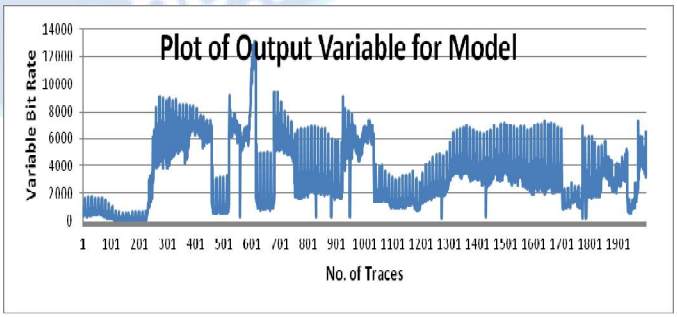

Fig. 2(b) : Plot of Output Variables used for Model Development

4.2 Generation of Train and Test Data

In this study, 2000 data set were used for training and 1000 data set were used for testing the network respectively. Normally, the data set needs to be divided into three parts. The first part is for the training, the second part for validation and the third part for testing. However, because the length of our sample data was not very big, we considered only two parts: training and testing. The only difference between a testing phase and a validation phase is that if the error rate of the validation phase increases, then the training stops. In this study, those two terms are used synonymously.

5 Model Develoment

5.1 Model Selection

In the present work Network Structure model consisting of one input layer with five input variables and an output layer consisting of variable bit rate of traffic data as the output variable. Here predictions were based on taking *N* previous patterns to predict one following pattern.

In the present work various ahead prediction models have been tested using the ICA. This includes 1- step ahead, 3-step ahead, 5-step ahead and 10-step ahead prediction models.

All the models include the time series predictions analysis, with the following modelling variables,

Table 1. Modelling Variables

	Input Variables (Time series VBR traces)					Output Variable
Model-1	VBR (t-4)	VBR (t-3)	VBR (t-2)	VBR (t-1)	VBR (t)	VBR(t +1)
Model-3	VBR (t-4)	VBR (t-3)	VBR (t-2)	VBR (t-1)	VBR (t)	VBR(t +3)
Model-5	VBR (t-4)	VBR (t-3)	VBR (t-2)	VBR (t-1)	VBR (t)	VBR(t +5)
Model-10	VBR (t-4)	VBR (t-3)	VBR (t-2)	VBR (t-1)	VBR (t)	VBR(t +10)

5.2 Parameter Selection

It is also important to select proper parameters for learning and refining process, including the initial step size (ss). In the present work commonly used rule extraction method i.e. subtractive clustering has been applied for identification and refinement. The model is simulated using the MATLAB version R2013a. In ICA Model using Tensor Analysis, the initial parameters of the model are identified using the subtractive clustering method. However, the parameters of the subtractive clustering algorithm still need to be specified. The clustering radius is the most important parameter in the subtractive clustering algorithm and is optimally determined through a trial and error procedure. By varying the clustering radius r_a between 0.1 and 1 with a step size of 0.01, the optimal parameters are sought by minimizing the root mean squared error obtained on a representative validation set. Clustering radius r_b is selected as 1.5 r_a. Default values are used for other parameters in the subtractive clustering algorithm [27].

Table 2 summerizes the results of types and values of model parameters used for training the model.

Table. 2. Types and values of parameters used for training model for 1-Step Ahead Model

No. of Nodes	44
No. of linear parameters	18
No. of non-linear parameters	30
Total no. Of parameters	48
No. of training data pairs	2000
No. of testing data pairs	1000
No. of rules generated	3

Figure 3 and 4 shows the comparative plots of observed and predicted Video Traffic Network Response data Prediction both for training and testing phases. The figures wisely demonstrate that (1) the model performance are in general accurate, where all data points roughly fall onto the line of agreement; (2) model using subtractive clustering is consistently superior in training phase than in testing phase.

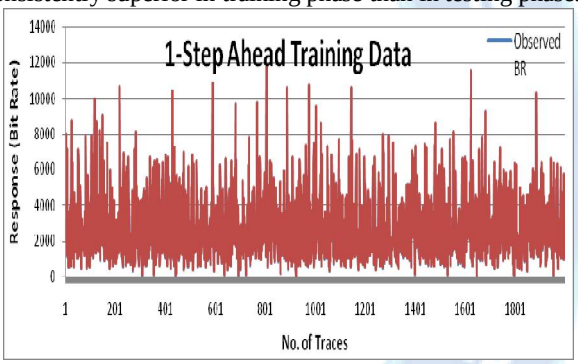

Fig. 3. Comparative plot of Actual and Predicted Video Traffic Network Response for Training Datasets

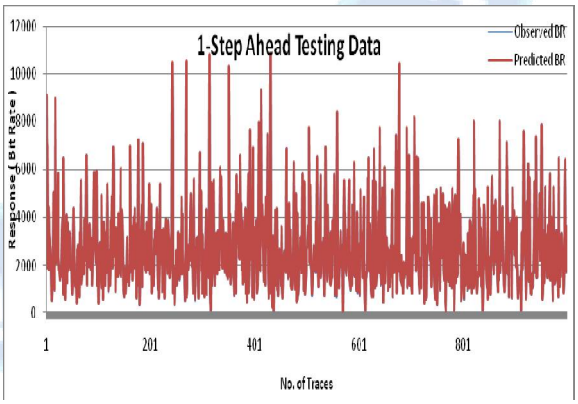

Fig. 4. Comparative plot of Actual and Predicted Video Traffic Network Response for Testing Datasets

6. Results and Discussions:

Model having five input variables are trained and tested by inference method and their performances compared and evaluated based on training and testing data. The best fit

model structure is determined according to criteria of performance evaluation. The performances of the models for various step ahead predictions are shown in Fig. 5 and their best MAE and RMSE values based on radius of influence r=0.50, both for training and testing data are given in Table 3 below.

Table 3. RMSE and MAE Values for Datasets

Prediction	MAE		RMSE	
	Training	Testing	Training	Testing
One Step Ahead	0.004938	0.00507	9.79	9.93
Three Step Ahead	0.0053	0.005537	12.6	12.44
Five Step Ahead	0.006282	0.006114	13.81	13.66
Ten Step Ahead	0.004296	0.004276	11.3	11.12

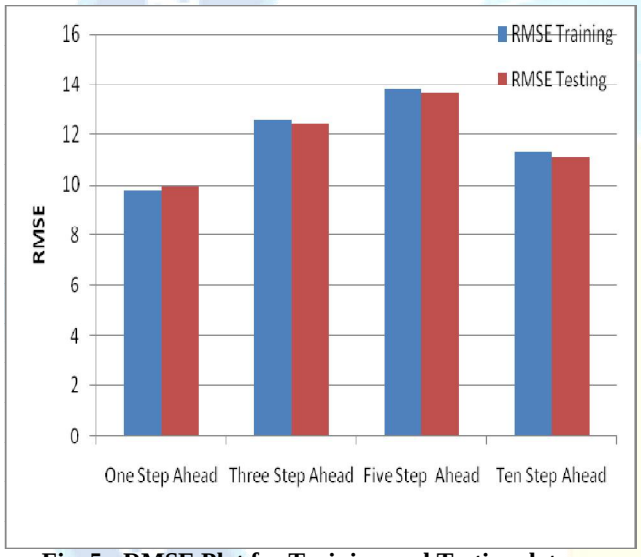

Fig. 5 : RMSE Plot for Training and Testing data

Further from the perusal of the scatter plot given in fig. 6 and 7 below it is evident that there is a good correlation between the predicted and the observed network traffic response output, as depicted by more or less straight trend line, with cluster of values at the beginning in case of training and testing data.

From the analysis of the above results, it is seen that the network traffic response prediction model developed using ICA technique has been able to perform well. The comparative analysis of all the four models show that the ICA has been able to best predict the one step ahead prediction, as compared to 3, 5 and 10-step ahead prediction models. But further analysis shows that 10-step ahead model has shown improvement in terms of RMSE and MAE with respect to 3 and 5-step ahead models. Also from the perusal of the charts given in Fig. 5, it is seen that in all the models the RMSE values for training and testing datasets are almost same, which again shows that the model has not overtrained.

Fig. 6. Scatter Plot of predicted versus observed Network Traffic Response data

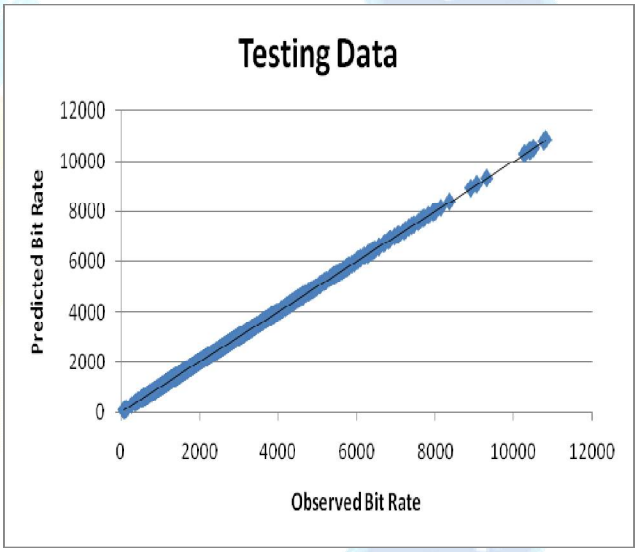

Fig. 7. Scatter Plot of predicted versus observed Network Traffic Response data

Further from the analysis of the observed and predicted values and as also from the corresponding graph given in fig. 3 & 4 above, both for training and testing data sets, it is clear that the ICA model has been able to perform better for both training and testing datasets. From the perusal of Fig. 6 and fig. 7 it is evident that all the data points corresponding to observed and predicted error values fall on the same line which again shows an excellent output.

7. Conclusion

In the present section, applicability and capability of Independent Component Analysis using CPD for time series analysis of network traffic response data prediction has been investigated. It is seen that the models are very robust, characterised by fast computation, capable of handling the noisy and approximate data that are typical of data used here

for the present study. Due to the presence of non-linearity in the data, it is an efficient quantitative tool to predict defect in softwares. The studies has been carried out using MATLAB simulation environment. The data used for the training and test sets is taken from the web log files of Telecommunication Networks Group, Technical University of Berlin, Germany [3]. Both the training and the test set consist of 3000 patterns (first 2000 representing the training, next 1000 the test set).

In case of 1-step ahead model, the RMSE value obtained from predictive model is 9.79 and 9.93 for training and testing datasets. Also, the MAE values obtained are 0.0049 and 0.0050 respectively, which clearly shows that MAE for training and testing dataset are almost similar. Further from Fig. 4.6 & 4.7 it is seen that computational model line almost closely follows the observed line. This again depicts the predictive superiority of the technique using tensor analysis of the data

References

[1] Evrim Acar, Daniel M. Dunlavy, and Tamara G. Kolda. 2009. Link prediction on evolving data using matrix and tensor factorizations. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops (ICDMW'09)*. 262–269. DOI:http://dx.doi.org/10.1109/ICDMW.2009.54

[2] Evrim Acar, Daniel M. Dunlavy, and Tamara G. Kolda. 2011. A scalable optimization approach for fitting canonical tensor decompositions. *Journal of Chemometrics* 25, 2 (2011), 67–86. DOI:http://dx.doi. org/10.1002/cem.1335

[3] Evrim Acar, Evangelos E. Papalexakis, Morten A. Rasmussen, Anders J. Lawaetz, Mathias Nilsson, and Rasmus Bro. 2014. Structure-revealing data fusion. *BMC Bioinformatics* 15, 1 (2014), 239. DOI:http://dx.doi.org/10.1186/1471-2105-15-239

[4] Rakesh Agrawal, Behzad Golshan, and Evangelos Papalexakis. 2015a. A study of distinctiveness in web results of two search engines. In *24th International Conference on World Wide Web, Web Science Track*. ACM. DOI:http://dx.doi.org/10.1145/2740908.2743060

[5] Rakesh Agrawal, Behzad Golshan, and Evangelos Papalexakis. 2015b. Whither social networks for web search? In *21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Sydney, Australia. DOI:http://dx.doi.org/10.1145/2783258.2788571

[6] Brett W. Bader and Tamara G. Kolda. 2007. Efficient MATLAB computations with sparse and factored tensors. *SIAM Journal on Scientific Computing* 30, 1 (Dec. 2007), 205–231.

[7] BrettW. Bader and Tamara G. Kolda. 2015.MATLAB Tensor Toolbox Version 2.6. Available online. (February 2015). http://www.sandia.gov/~tgkolda/TensorToolbox/.

[8] Jonas Ballani, Lars Grasedyck, and Melanie Kluge. 2013. Black box approximation of tensors in hierarchical Tucker format. *Linear Algebra and Its Applications* 438, 2 (2013), 639–657. DOI:http://dx.doi.org/10.1016/j.laa.2011.08.010

[9] Rasmus Bro and Henk A. L. Kiers. 2003. A new efficient method for determining the number of components in PARAFAC models. *Journal of Chemometrics* 17, 5 (2003), 274–286. DOI:http:// dx.doi.org/10.1002/cem.801

[10] Kai-Wei Chang, Wen-tau Yih, Bishan Yang, and Christopher Meek. 2014. Typed tensor decomposition of knowledge bases for relation extraction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1568–1579. DOI:http://dx.doi.org/10.3115/v1/d14-1165

[11] Jeremy E. Cohen, Rodrigo Cabral Farias, and Pierre Comon. 2015. Fast decomposition of large nonnegative tensors. *IEEE Signal Processing Letters* 22, 7 (2015), 862–866. DOI:http://dx.doi. org/10.1109/lsp.2014.2374838.

[12] Acar, E., C_ amtepe, S. A., and Yener, B. 2006. Collective sampling and analysis of high order tensors for chatroom communications. In ISI 2006: Proceedings of the IEEE International Conference on Intelligence and Security Informatics. Lecture Notes in Computer Science, vol. 3975. Springer, 213{224.

[13] Acar, E., Kolda, T. G., and Dunlavy, D. M. 2009. An optimization approach for fitting canonical tensor decompositions. Tech. rep., Sandia National Laboratories, Albuquerque, New Mexico and Livermore, California. Feb. Submitted for publication.

[14] Acar, E. and Yener, B. 2009. Unsupervised multiway data analysis: A literature survey. IEEE Transactions on Knowledge and Data Engineering 21, 1 (Jan.), 6{20.

[15] Hasan, M. A., Chaoji, V., Salem, S., and Zaki, M. 2006. Link prediction using supervised learning. In Proc. SIAM Data Mining Workshop on Link Analysis, Counterterrorism, and Security.

[16] Kolda, T. G. and Bader, B. W. 2009. Tensor decompositions and applications. SIAM Review 51, 3 (Sept.), 455{500.

[17] Sharan, U. and Neville, J. 2008. Temporal-relational classi_ers for prediction in evolving domains. In ICDM '08: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. IEEE Computer Society, 540{549.

[18] Ashish Jain, Rajeev Sharma, Gireesh Dixit, Varsha Tomar ," Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages", 2013 IEEE International Conference on Communication Systems and Network Technologies.

[19] Luo Q.;"Advancing Knowledge Discovery and Data Mining",First International Workshop on Knowledge Discovery and Data Mining, 2008.

[20] B. W. Bader and T. G. Kolda, *Algorithm* 862: *MATLAB tensor classes for fast algorithm prototyping*, ACM Trans. Math. Software, 32 (2006), pp. 635–653.

[21] P. M. Kroonenberg, *Three-Mode Principal Component Analysis: Theory and Applications*, DWO Press, Leiden, The Netherlands, 1983.