Report – Short Term Mobility Program

Time-varying Machine Learning for Big-Data

Ercan Engin Kuruoglu

Fraunhofer Heinrich Hertz Institute, Machine Learning Laboratory, Berlin, Germany

Over the last decade, various fields of science and engineering have been experiencing a change of paradigm with the increasing availability of vast amounts of data. This data explosion commonly referred to as "Big Data" has been made possible due to the developments of sensing apparatus, increasing computational power, data storage facilities getting cheaper due to the advances in electronic technology and the increasing ease of sharing data. Examples include genomic data, due to the cheaper sequencing technology of various species such as homo sapiens, primates, yeast, agricultural plants, and cancer genome, astrophysical data such as the full sky maps of stars and galaxies or the cosmic microwave radiation obtained by recent satellite missions, neurological data, seismic data, environmental sensing data available from vast amounts of sensors around the world and social networks data in the digital realm.

These vast amounts of data place new challenges, most importantly on how to derive required information out of an ocean of data samples. In particular, the heterogeneous nature of data pressure us to leave our system models which vision data as output of some input to the system. Moving to multiple input multiple output (MIMO) models are not a cure either due to the loss of track of causality in the modern data realm. Instead of causality relations now it is more pertinent to understand relations between different variables. Again due to the nature of data which is noisy, we are forced to do this statistically, in terms of statistical relations.

A full analysis of available data requires the building of models of statistical dependencies between each pair of observed variables. This leads to the building of networks. Two important challenges in building statistical networks are to avoid duplication of statistical dependence information and keeping the network at a manageable dimension. The first requires working with conditional dependencies and the latter requires enforcing sparsity. Various work are available in the literature that have the objective to provide solutions to the network modelling problem. There are important restrictions, however, how this problem has been handled in previous work. Firstly in most work the data is assumed to be "normal," or Gaussian that is only first and second order statistics (correlations) are considered for modelling dependence between data variables. This is very approximate in the sense that it assumes that correlations capture all statistical dependence between variables. This is the case only for "normal" or Gaussian distributed data. To have a general framework valid for any distributions, mutual information rather than statistical correlations should be considered. This can lead to more successful models for data which are skewed and containing rare events.

The second and more important restriction of existing work is the assumption of data being stationary or even constant. In various application areas, such as signal processing, computational biology, finance, seismology, climatology, telecommunications, however, the data is in the form temporal sequences and evolve over time. In most of the cases, it is even hard to claim that the statistical properties are stationary over time and one can at best assume piecewise stationarity. Expanding more on the applications, in the case of genomic data as an organism is born, and becomes an adult the gene activation levels and their relation to each other changes, similarly in the case of circadian cycles or in the development of cancer cell we see the dynamics of gene expression change over time. In the case of seismology data, the seismic displacement at different locations which are related via detailed net of seismic fault lines change over time. Similarly in climatology, weather conditions (precipitation, temperature, pressure) can be seen as a

Kuru

complicated network of interactions which change over time. Finally, in brain neural networks, as evidenced with various experiments with fMRI, the connections between different parts of brain change over time as reaction to stimulus.

Our aim has been to develop new methodologies for modelling multivariate data of big dimensions which change over time. In contrast to the existing models in the literature the proposed models will be stochastic and hence with potential for future predictions.

The general mathematical framework we propose is Bayesian Networks which are established tools for efficiently modelling multivariate data. A DBN (Dynamic Bayesian Network) is an extension of Bayesian network to temporal domain. In time domain, conditional dependencies can be modelled between random processes as opposed to simply random variables within as well as across time epochs. The conditional distributions in DBN are assumed to be homogeneous, i.e. the structure and parameters of a network are maintained constant throughout the time. Taking this into account, a DBN is simply constructed by unwrapping a Bayesian network in time domain that causes significant simplification to the model learning procedure. At the same time, this assumption constrains the strength of DBN in modelling non-stationary sequences, where intrinsic relationships between different variables change in time. These non-stationary sequences are present everywhere in the nature, for example the gene interactions in different stages of a life cycle. Obviously, usage of a stationary statistical model is insufficient for modelling gene expression data sequences at all time instances.

Learning a time varying network is not a trivial problem. One may try naively to learn a dynamically changing network independently for each time epoch. However, this is a complex task as there are very little available observations at one time epoch for most applications in real life. One way to overcome the problem of data scarcity is to divide temporal sequences into segments – stationary epochs, with an assumption that in each epoch data are generated from the same probability distribution. However, the lack of knowledge about models in each segment makes the problem more complex. Moreover, the solution space grows exponentially with the length of time sequence. From another point of view, since observations are most often distorted by noise, statistics can be recovered from these modifications of signal.

To overcome all the difficulties mentioned above, one of the propositions is to expand DBN to nonstationary scenarios by introducing various additional conditions on the type of a network and how the network can change in time. The works done before have been mainly concentrated on nonstationary models with static structure. One of the most popular models is the time-varying autoregression model (TVAR). TVAR is able to describe nonstationary linear dynamic systems, coefficients and noise variances, which continuously change with time. In order to estimate recursively the regression parameters, normalized least squares algorithm can be used. And an error of estimation is shown to be bounded when the model parameters change smoothly.

The methodology that we propose to learn dynamical Bayesian networks is sequential Monte Carlo (or particle filtering). Sequential Monte Carlo iteratively solves a stochastic filtering problem where the hidden variables are estimated using a priori information and observable quantities. Sequential Monte Carlo has reached important level of success in tracking problems in computer vision and radar signal processing. It has several advantages to rival models: 1. It is general, it does not assume linearity or Gaussianity. It is valid also for nonlinear systems/interactions and non-Gaussian populations. 2. It incorporates prior information about variables in the form of prior pdfs which allows one to use the method adaptively with the arrival of new data. Old posteriors become new priors. It is hence a seamless learning method. 3. It is specifically easy to develop for regression models. 4. As new data samples arrive, it provides incremental updates to the model avoiding from scratch calculation unlike most other Bayesian methods or machine learning methods. 5. Recently, it has also been extended to multi-object tracking and therefore has a ready mathematical framework.

We used this mathematical framework in tracking the statistical dependencies between various variables. The observables are the absolute values of the variables that is the node values in the network while the algorithm learns the branch values, that is, the correlations or statistical dependences between variables. We have made an initial implementation on genetic data [3] and obtained very promising results.

There are several ways this approach needs to be developed and during my visit in the Fraunhofer-Heinrich Hertz Institute, we made progress on two aspects.

1. Network Structure Optimization:

The main problem with network modelling using stochastic models is the explosion of model volume (geometric increase) with the increasing number of variables. We need new methods that predict before hand the important variables in the network and concentrate only on them. This cannot be a simple preprocessing step since due to the time varying nature of the data, new variables can gain importance and some can lose importance. The new method should be able to adapt to this and allow on and off data models. This issue is an open problem and requires a concentrated research effort carrying some of the existing (but unknown to other fields) methods from mathematics, jump Markov processes in particular.

In addition to the problem of deciding which nodes are important in a network, it is also important to decide which edges or branches are important in a network. While the nodes describe the variables, the edges or brances describe the relations between them. It is important to consider these two problems in combination rather than separately. The combined problem is in general non-convex and therefore steepest descent type methods do not lead to global optimal solutions.

We have developed two approaches to solve the joint problem of node and edge optimization. The first of the methods does random walks starting from a sub-optimal solution. The random walks are defined on the structure of the network. Better solutions are accepted immediately and worse solutions are accepted according to some Boltzmann statistics allowing the algorithm to do hill-climbing and avoid local minima.

The second method makes a global Bayesian modelling of the nodes and the edges by assigning a multivariate distribution and then executing random walks with the aim of finding maximum a posterior solution.

2. Network Compression:

Even when the important variables are detected, fully connected networks are expensive, we should develop efficient thresholding methods that remove unimportant connections. Sparsity enforcing priors are a common approach but generally this is done without much conscience into the assumptions these priors embody. We propose an alternative approach based on Information Theory of Shannon, in particular the Rate-Distortion theory. Using Rate-Distortion theory, we can compress networks controlling the loss of information. Various other methods are possible for compressing networks; however, the rate-distortion theory provides a well-established theory that gives the estimate of distortion in the modelling performance of the network due to compression before hand. One can hence find the optimal network given the level of distortion one can tolerate in a disciplined way.

Kuruo

3. Applications:

The methodology which we started to develop has two main application areas. Firstly for modelling interaction networks. Interaction networks are important modelling devices for applications such as gene-interaction networks, protein-protein networks, seismology networks, meteorological networks, financial networks, brain networks, etc. In the future, we would like to apply the methodology on these applications.

The other main avenue for applications is deep learning neural networks. The above described methodologies can be applied also for compression of deep neural networks on which my host Dr Samek has important work and also on the optimization of the structure. As first step we have worked on stationary networks, and now we are looking into time-varying networks.

As future work, we will extend Dr Samek's work on rate distortion based compression of deep learning neural networks to the time varying deep learning neural networks.

4. Publications:

Once the works are complete, we plan to make international journal publications on:
a. "Deep Neural Network structure optimization by Bayesian Methods" in IEEE Transactions on Neural Networks and Learning
b. "Deep Neural Network structure optimization by Annealing Methods" in IEEE Transactions on Neural Networks and Learning
c. "Interaction Network Compression using Rate-Distortion Theory" in IEEE Transactions on Signal Processing on Graphs

5. Other activities during STM

A seminar was given in Fraunhofer-Heinrich Hertz Institute with title "Is the Gaussian Distribution Normal? Data Analysis with alpha-Stable Distributions". Motivated by this seminar, a PhD student from the Machine Learning Laboratory has started working on image synthesis using Levy flights.

Potsdam Institute of Climate Change was visited as the guest of the Director of the Institute, Prof Kurths and a seminar was given with title Is the Gaussian Distribution Normal? Data Analysis with alpha-Stable Distributions". Further meetings were held on time varying network modelling and a number of researchers were motivated to collaborate on the application of time varying network modelling to climatology problems.