MODELLING OF STATISTICAL FLUCTUATIONS OF INFORMATION FLOWS BY MIXTURES OF GAMMA DISTRIBUTIONS

Andrey Gorshenin
Institute of Informatics Problems,
Russian Academy of Sciences
Vavilova str., 44-2, Moscow, Russia
Email: agorshenin@ipiran.ru

Victor Korolev
Moscow State University
Leninskie Gory, Moscow, Russia
Institute of Informatics Problems,
Russian Academy of Sciences
Email: bruce27@yandex.ru

KEYWORDS

Information flows modelling; Stochastic structure; Probability mixtures

ABSTRACT

The paper describes statistical approach to the analysis of traffic of information flows. Stochastic structure of traffic process can be modelled by finite probability mixtures, e.g., mixtures of gamma distributions. The approach is demonstrated on real data from the official website of the Russian Academy of Sciences.

INTRODUCTION

Developing methods of investigating of probabilistic and statistical regularities related to rare events is an important area in the modern theory of probability. In particular, the Poisson theorem is the basis for constructing mathematical models in telecommunication systems. Classical stochastic models of telecommunication systems are based on the hypothesis that the data flows are Poisson. The assumption of the Poisson character of flow entails the fact that the development of a random process in future does not depend on its past and is determined only by its value at the current time. But this model is ideal, because real processes in the telecommunication systems do not satisfy the ideal conditions that imply Poissonity (Gnedenko, Korolev, 1996). So, to describe real complex information systems some generalization is needed.

GAMMA MIXTURE MODEL FOR INFORMATION FLOWS

In general, it may be assumed that the flow of events related to traffic is chaotic. The entropy reasoning leads us to the conclusion that the best model for the distribution of inter-arrival times in a completely chaotic flow is the exponential distribution resulting in that the flow itself is Poisson. The telecommunication systems are not closed systems. Therefore, it may be assumed that the exponential model can be regarded as conditional after conditioning with respect to information flows.

Probably, the best model of homogeneous chaotic stochastic flow is the Poisson process. But because of heterogeneity of chaos in real information systems, compound Cox process (Korolev, 2011) should be used instead of Poisson process. So, we have special reasons to examine finite gamma mixtures for modelling information flows.

For the investigation of the fine structure of information flows we assume the total sample to be locally homogeneous and suggest that within the window (number of elements) the sample is homogeneous. Then the window moves in the direction of the astronomic time making it possible to trace the evolution of the parameters of the gamma distribution in time. This idea is the essence of a method which is called "moving separation of mixtures" (MSM method). Accordingly, the original sample is split into smallers subsamples (windows), and the system is analysed within each window. The MSM method allows to observe time evolution of components. We can reveal dominating components, which form the process, and noise components which appear due to computational inaccuracy.

EM-ALGORITHM FOR MIXTURES OF GAMMA DISTRIBUTIONS

To estimate the parameters of gamma mixtures, the maximum likelihood method is used realized by the EM-algorithm. The formulas for the calculation of the estimates of the parameters on iterative steps have the following form. The shape parameter can be found by a numerical solution of the equation

$$\log r_i^{(m)} - \psi(r_i^{(m)}) = -\frac{\sum\limits_{j=1}^{n} g_{ij}^{(m)} \log \frac{x_j}{A_i^{(m)}}}{\sum\limits_{j=1}^{n} g_{ij}^{(m)}}, \qquad (1)$$

Proceedings 27th European Conference on Modelling and Simulation ©ECMS Webjørn Rekdalsbakken, Robin T. Bye, Houxiang Zhang (Editors)
ISBN: 978-0-9564944-6-7 / ISBN: 978-0-9564944-7-4 (CD)

where $\psi(\cdot)$ is the digamma function, the quantity $A_i^{(m)}$ has the form

$$A_i^{(m)} = \frac{\sum\limits_{j=1}^{n} x_j g_{ij}^{(m)}}{\sum\limits_{j=1}^{n} g_{ij}^{(m)}}, \tag{2}$$

$$g_{ij}^{(m)} = \frac{p_i^{(m)} f_{r_i^{(m)}, \theta_i^{(m)}}(x_j)}{\sum\limits_{l=1}^{k} p_l^{(m)} f_{r_l^{(m)}, \theta_l^{(m)}}(x_j)}, \tag{3}$$

on each step the scale parameter is determined by the relation

$$\theta_i^{(m)} = \frac{A_i^{(m)}}{r_i^{(m)}}, \; i = 1, \ldots, k, \tag{4}$$

and the weights are given by the formula

$$p_i^{(m+1)} = \frac{1}{n} \sum_{j=1}^{n} g_{ij}^{(m)}. \tag{5}$$

APPLICATION FOR REAL DATA

The following section deals with the application of the method for real information data.

Let us consider the average time that the visitors spent on website (see Fig. 1). The quantity equals the difference between time of last and first page browsing during the visit.

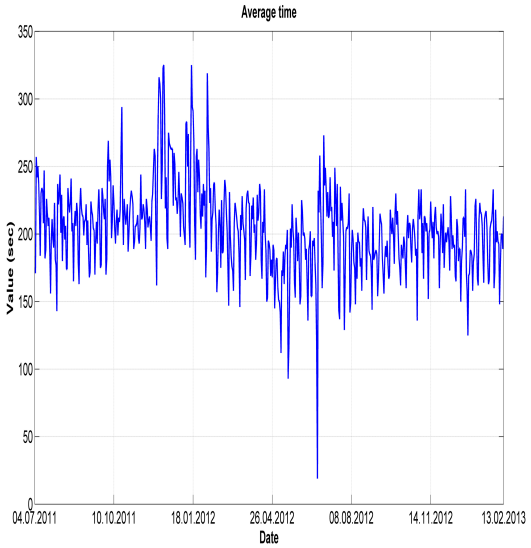

Figure 1: Average time of page browsing

The basic assumption is the existence of 3 components in the mixture. The EM-algorithm is applied in the moving mode, the window width is equal to 200 elements (see Fig. 2).

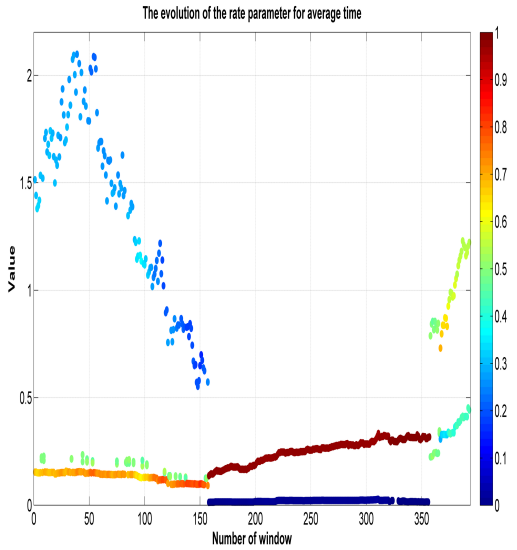

Figure 2: The evolution of rate parameter of gamma distribution in mixture

On the abscissa, the number of the current window in MSM method is plotted. On the ordinate, the values of gamma distributions parameter estimates are located (for the rate parameter of gamma distribution). The colormap corresponds to the weights of components. We can see 2 components during the whole period under review.

Let us consider the number of unique users with at least one visit on website during the period under review (see Fig. 3).

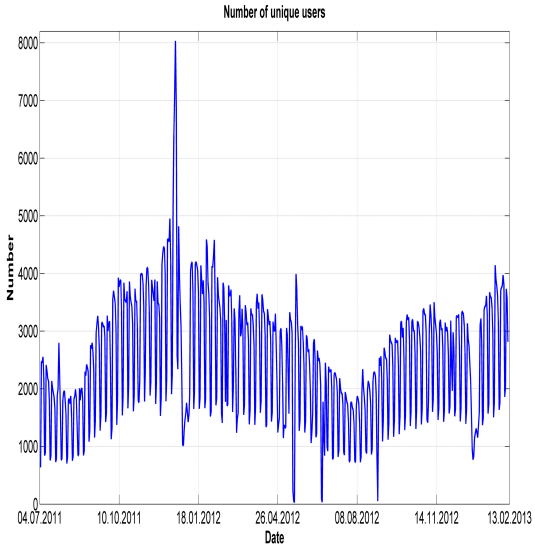

Figure 3: Number of unique users

The basic assumption is the existence of 4 components in the mixture. The EM-algorithm is ap-

plied in the moving mode, the window width is equal to 200 elements (see Fig. 4).

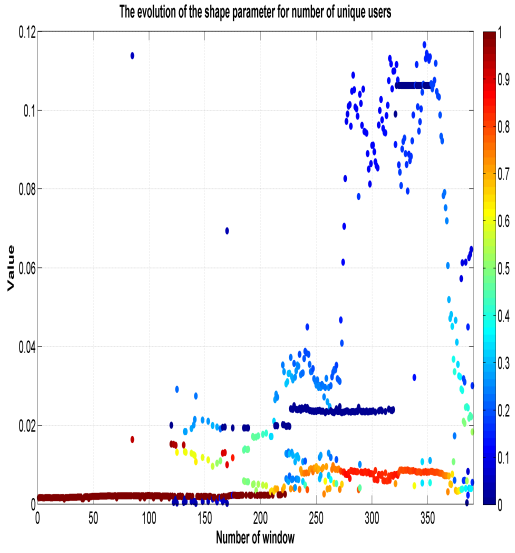

Figure 4: The evolution of the shape parameter of gamma distribution in mixture

On the abscissa, the number of the current window in MSM method is plotted as before. On the ordinate, the values of gamma distributions parameter estimates are located (for the shape parameter of gamma distribution). The colormap corresponds to the weights of components. We can see $2-3$ different components during the period under review. One component exists through the whole time period, but another components arise at some positions and disappear from our view when window is moved.

Naturally, the essence of the components for each considered data type may be different. But the components usually can be interpreted according to knowledge domains. It is one of the possible ways for analyzing and forecasting fine structure of data.

CONCLUSIONS

The representation of the probability distribution of traffic in information flows as a finite mixture of gamma-distributions makes it possible to reveal the stochastic structure of the information flow, that is, separate a finite number of typical sub-structures or classes of more or less similar (within a type) claims (or jobs) being executed at a time, determine the proportion of each substructure and trace the evolution of the stochastic structure in time. As this is so, the number of typical classes and their proportions can change in time also. The parameters of gamma distribution can be interpreted. For example, the ratio of these parameters determines the mean value of each class of jobs or the intensity of the sub-flow of jobs of the specific type. We obtained several different components in each of considered situations which could characterize the behavior of structures forming the information flows. The analysis of the components of flows gives an opportunity to create more detailed models of the functioning of information systems.

The efficiency of the methodology can be applied to various information systems. For example, the results of modeling of information flows in financial markets (the evolution of limit order books) can be found in (Gorshenin et al., 2012).

The new method of automatic data classification is represented in paper. The MSM methodology for traffic research is used for the first time. This approach can be used as basis of investigation in very important modern practical field – big data analysis. The big data in information technologies is combination data with innovation methods of analysis while traditional approaches are not inapplicable. The most important problem of the field is creation of strategies of working with the specific large data. The results mentioned above imply that described approach represents one of the perspective technique for exploring efficiency of software functioning in a changing computing environment. The main advantage of the MSM approach is combination of statistical analysis and data mining methods. Leaving out computational efficiency problem, in the paper we demonstrated reasonableness of approach in the particular example of data with modest size.

The authors express their gratitude to A.V. Bosov and service `http://metrika.yandex.ru/` for the permission to use concrete data.

The research is supported by the Russian Foundation for Basic Research (projects 12-07-00115a, 12-07-31267mol_a and 11-07-00112a).

REFERENCES

Gnedenko B. V., Korolev, V. Yu. 1996. "Random Summation: Limit Theorems and Applications." Boca Raton: CRC Press, 41-152.

Gorshenin A., Doynikov A., Korolev V. and Kuzmin V. 2012. "Statistical Properties of the Dynamics of Order Books: Empirical Results." *VI International Workshop "Applied Problems in Theory of Probabilities and Mathematical Statistics Related to Modeling of Information Systems."* M.: Institute of Informatics Problems, RAS, 2012, 31-51.

Korolev, V. Yu. 2011. "Probabilistic and Statistical Methods of Decomposition of Volatility of Chaotic Processes." Moscow: Moscow University Publishing House (in Russian), 31-76.

AUTHOR BIOGRAPHIES

ANDREY GORSHENIN is Candidate of Science (PhD) in physics and mathematics, senior scientist, Institute of Informatics Problems, Russian Academy of Sciences.

VICTOR KOROLEV is Doctor of Science in physics and mathematics, professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M.V. Lomonosov Moscow State University; leading scientist, Institute of Informatics Problems, Russian Academy of Sciences.