Constructing a Neural-Net Model of Network Traffic Using the Topologic Analysis of Its Time Series Complexity

N. Gabdrakhmanova(✉)

People's Friendship University of Russia, Moscow, Russia
gabd-nelli@yandex.ru

Abstract. The dynamics of data traffic intensity is examined using traffic measurements at the interface switch input. The wish to prevent failures of trunk line equipment and take the full advantage of network resources makes it necessary to be able to predict the network usage. The research tackles the problem of building a predicting neural-net model of the time sequence of network traffic.

Topological data analysis methods are used for data preprocessing. Nonlinear dynamics algorithms are used to choose the neural net architecture. Topological data analysis methods allow the computation of time sequence invariants. The probability function for random field maxima cannot be described analytically. However, computational topology algorithms make it possible to approximate this function using the expected value of Euler's characteristic defined over a set of peaks. The expected values of Euler's characteristic are found by constructing persistence diagrams and computing barcode lengths. A solution of the problem with the help of R-based libraries is given. The computation of Euler's characteristics allows us to divide the whole data set into several uniform subsets. Predicting neural-net models are built for each of such subsets. Whitney and Takens theorems are used for determining the architecture of the sought-for neural net model. According to these theorems, the associative properties of a mathematical model depend on how accurate the dimensionality of the dynamic system is defined. The sub-problem is solved using nonlinear dynamics algorithms and calculating the correlation integral. The goal of the research is to provide ways to secure the effective transmission of data packets.

Keywords: Computational topology · Persistence · Stability · Neural network

1 Introduction

The topicality of the study is determined by the following reasons. The continuing development of telecommunication and Internet services sets new requirements for the bandwidth of telecommunication channels. The presence of a great deal of various services in a single physical transmission medium at pick hours can bring about the overloading of switching and routing devices in trunk lines and, as result, a reduction of many services. The wish to prevent failures of trunk line equipment and take the full advantage of network resources makes the problem of effective use of the

B. Kryzhanovsky et al. (eds.), *Advances in Neural Computation, Machine Learning, and Cognitive Research*, Studies in Computational Intelligence 736,
DOI 10.1007/978-3-319-66604-4_14

telecommunications channel bandwidth very important (the direct widening of the bandwidth inevitably leads to an increase of service costs). It is necessary to have effective traffic control methods that could use statistical data to predict the traffic intensity. A lot of modern publications deal with mathematical models of different types of network traffic [1–3]. The complexity and relevance of this problem urge further research in the field.

2 The Topological Data Analysis

The topological data analysis is a new theoretical trend in the field of data analysis. The approach allows the determination of topological data structures. Recent advancements in the field of computational topology make it possible to find topological invariants in data collections [2, 4, 5].

The point of the analysis is that stable properties are to be immune to noise, distortions, errors, lack of data. The practice of using the analysis in different fields shows that the supposition is true and stable topological properties can provide a lot of information about data collections. Persistence diagrams are one of basic tools of computational topology. They make it possible to get useful information about excursion sets of a function. Below are the basic definitions (according to [4]).

Let X be a topological space being triangulated, f be a continuous tame function defined over space X. Let us introduce the notation $U_a = f^{-1}(-\infty, a]$ for $a \in R$. When moving upwards, components U_a can merge or produce new components. It is possible to trace how the sub-level topology changes with a by examining homologies of these sets with, say, persistence homologies. Parameter $a \in R$ is called the homological critical value if for certain k the homomorphism induced by nesting $f_* : H_k(U_{a-\varepsilon}) \to H_k(U_{a+\varepsilon})$ is not an isomorphism for any sufficiently small $\varepsilon > 0$ (homology groups are considered with coefficients in Z_2). Continuous function f is called tame function if it has a finite number of homological critical values. When $b \leq a$, then $U_b \subseteq U_a$. Let us denote a set of connectivity components as $C(a) = C(U_a)$. It is possible to define a functional – Euler characteristic – over a set of sub-levels of U_a. Let $X \subset R^2$. Then, in the terms of algebraic topology, Euler's number is $\chi(U_a) = \beta_0 - \beta_1$, where β_0, β_1 are the ranks of the first two homology groups. Functional $\chi(U_a)$ measures the field topological complexity on the sub-level set. Note that for function f it is possible to deal with a set of super-levels $U_a = f^{-1}[a, \infty)$ instead of sub-levels.

Let us define the persistence diagram according to [5]. Let f: $X \to R$ be a tame function. Let $a_1 < a_2 < \ldots < a_n$ be critical homological values. Let us consider interjacent values $b_0, b_1, \ldots, b_n : b_{i-1} < a_i < b_i$. Let us supplement the chosen points in the following way: $b_{-1} = a_0 = -\infty$; $b_{n+1} = a_{n+1} = +\infty$. Let us define the multiplicity of point $(a_i < a_j)$ for each couple of indices $0 \leq i < j < n+1$ by setting $\mu_i^j = \beta_{b_{i-1}}^{b_j} - \beta_{b_i}^{b_j} + \beta_{b_i}^{b_{j-1}} - \beta_{b_{i-1}}^{b_{j-1}}$, where $\beta_x^y = \dim(\mathrm{Im}(f_x^y)), f_x^y : H_k(U_x) \to H_k(U_y)$. Persistence diagram $D(f) \subset R_2$ of function f stands for a set of points (a_i, a_j) $(i, j = 0, 1, \ldots, n+1)$ adjusted for multiplicity μ_i^j in combination with a set of diagonal points $\Delta = \{(x, x) | x \in R\}$ adjusted for infinite multiplicity.

The immunity of a persistence diagram to perturbations of function f is its remarkable feature. Persistence diagrams can be used to calculate the lengths of the barcodes of connectivity components. Here the term barcode stands for the component lifetime. Let us denote the summarized lengths of barcodes of two homology groups H_0 and H_1 as L_0 and L_1 correspondingly. Then the mean of the Euler characteristic can be determined [2] as

$$\chi = L_0 - L_1. \tag{1}$$

3 Setting the Problem

A second-level interfacial switch of a backbone line provider is taken as a test object in the paper. The traffic coming to each port of the switch is integrated traffic from user groups belonging to a particular region. The explanatory drawing is given in Fig. 1. The Cacti software (SNMP interface protocol) was used to gather statistic data. The information about the degree of network usage is more useful in practice. The knowledge of the number of packets in unit time can be misleading. For this reason the aggregate quantity $x(t)$ – traffic intensity (in bits) at moment t – is taken as an observable variable. The extension of data is 10080 points or 7 days. The plot of traffic intensity measured at port GE 0 is shown in Fig. 2. Each point in this plot represents a number of bits going through the trunk in one minute's time.

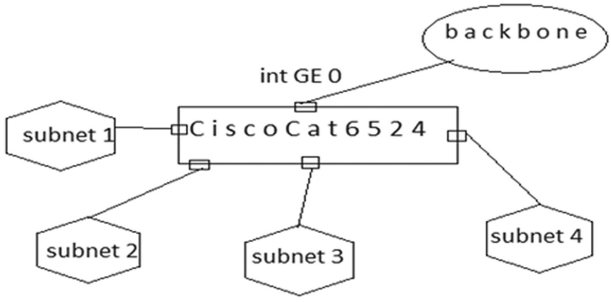

Fig. 1. The measurement arrangement.

So the goal is to construct a mathematical model for the m-step prediction of traffic intensity using observations $\{x(t), t = 1, 2, \ldots, N\}$, where N is the number of points. The estimates of Euler's characteristics are used here as indication of network usage. The following algorithm is proposed. The whole data collection is to be divided in clusters with different Euler's characteristics. A neural-net prediction model is to be built for each cluster using nonlinear dynamics methods. Below is the result of the experimentation.

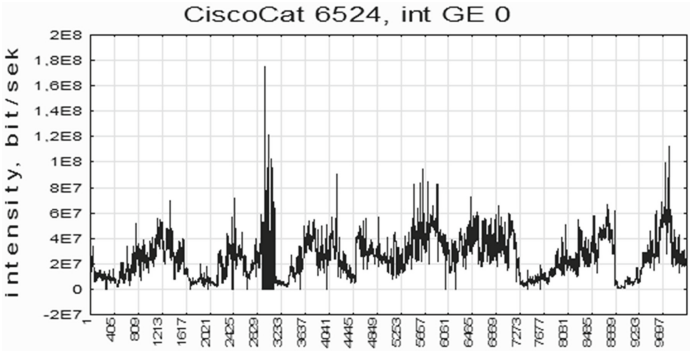

Fig. 2. The traffic intensity plot at port GE 0.

4 Topological Invariants Calculated for a Traffic Intensity Sequence

Packet TDA from a public repository of R packets was used as a library for finding stable homologies. The packet has a broad toolkit for topological data analysis by topological methods.

Before finding topological characteristics, the whole data collection was divided in some portions. Each portion held data acquired in two hours' time. For each portion persistence diagrams, barcodes were determined and Euler's characteristic estimates were calculated by formula (1).

The following algorithm was used to find estimates of Euler's characteristic in the TDA packet. A triangulation grid was first built using function Grid(). Then function gridDiag was used to produce matrix Diag. Function gridDiag evaluates the actual value of the function by the triangulation grid, generates simplex filtration using these values, and calculates constant homologies from the filtration. Figure 3 shows the persistence diagrams for one portion of data. The birth time of a component is plotted

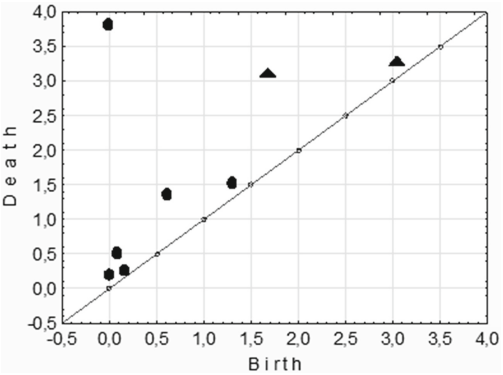

Fig. 3. The plot on the right shows the persistence diagram of the superlevel sets of the KDE.

as abscissas; the death time is plotted as ordinates. The dots correspond to zero-dimensional simplexes, the triangles mark single-dimensional simplexes. Figure 4 presents the barcode chart of zero-dimensional simplexes. Table 1 gives the estimates computed for different (n = 15) portions of the object. The following notation is used in the table: n is the number of a portion (interval), L0 and L_1 are the summarized barcode lengths of zero- and single-dimensional simplexes, χ is the estimate of Euler's characteristic (1). The plot in Fig. 5 shows Euler's characteristic as function of n. The horizontal axis represents the number of an interval and Euler's characteristic is measured on the vertical axis.

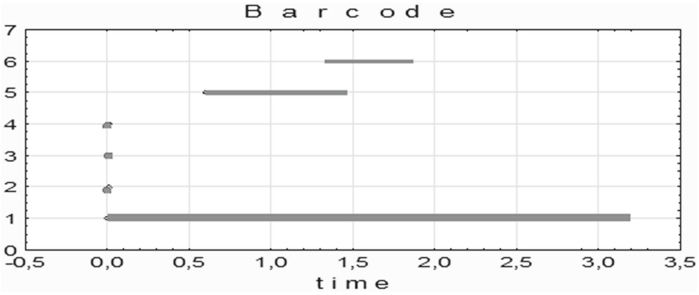

Fig. 4. Barcode

Table 1. The estimates of Euler's characteristic

n	1	2	3	4	5	6	7	8	9	10	11	12
L_0	3.7	4.1	3.6	3.7	2.2	1.7	2.0	2.0	1.8	2.4	3.4	3.6
L_1	2.3	1.6	1.7	1.8	2.5	1.7	2.0	2.0	1.5	3.3	1.8	2.3
χ	1.5	2.6	1.9	1.8	−0.3	−0.1	0.03	0.04	0.3	−0.8	1.5	1.4

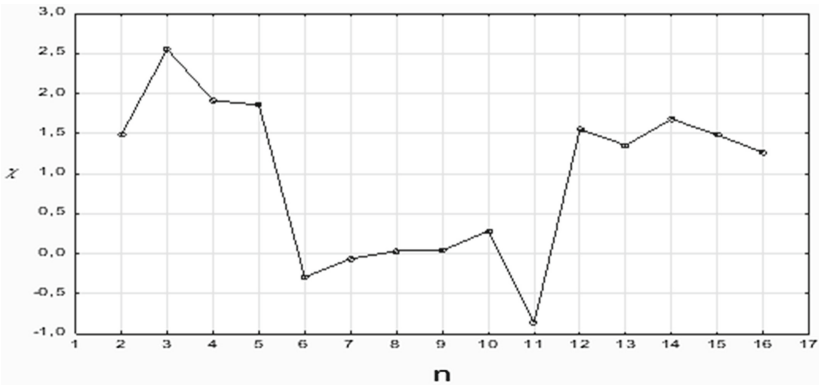

Fig. 5. Euler's characteristic as function of n.

The results prove that Euler's numbers are a stable characteristic of traffic intensity. At the next stage the portions with the same [χ] (where [.] is the integer of a number) are united in a single cluster. A neural-net prediction model is built for each cluster.

5 Building the Neural-Net Model of the Data

Methods of nonlinear dynamics are used to construct a neural-net model for a selected cluster. The subproblem is set as follows. Let $\{x(t)\}_{t=1}^{N}$ be measurements of a particular observable scalar component of a d_1-dimensional dynamic system $\bar{y}$. On the whole, the dimensionality and behavior of the dynamic system are not known. For a given time sequence it is necessary to build a model that would incorporate the dynamics responsible for the generation of observations $x(t)$. According to Takens' theorem, the geometrical structure of the dynamics of a multivariable system can be restored using observations $\{x(t)\}_{t=1}^{N}$ in a D-dimensional space built around new vector $\bar{z}(t) = \{x(t), x(t-1), \ldots, x(t-(D-1))\}^T$ (where $D \geq 2d_1 + 1$). The evolution of points $\bar{z}(t) \rightarrow \bar{z}(t+1)$ in the restored space corresponds to the evolution of points $\bar{y}(t) \rightarrow \bar{y}(t+1)$ in the initial space. The procedure of searching for a suitable D is called nesting. The least value of D at which the dynamic restoration is achieved is called the dimension of the nesting. The algorithm offered by P. Grassberger and I. Proccaccia in 1983 makes it possible to evaluate D using a time sequence.

After D is estimated, the problem at hand can be formulated in the following way. There is time series $\{x(t)\}_{t=1}^{N}$ and restoration parameters (D = 11 in our case) are set. For N_1 vectors $\bar{z}(t) = \{x(t), x(t-1), \ldots, x(t-(D-1))\}^T$ the values of the sought-for function $F(t) = F(\bar{z}(t))$ are known (because the terms of the time series following $\bar{z}(t)$ are known). It is necessary to find the value of the sought-for function at new point $\bar{z}(t)$, $\hat{x} = F(\bar{z})$.

Neural nets of the multiple-layer perceptron type [6] are used to tackle the problem. Only the key results are given below. Figure 6 shows the graph of traffic intensity on a set of test points. The horizontal axis represents time, the vertical axis shows the normalized traffic intensity; the solid line corresponds to experimental data x, the dashed line represents theoretical results $\hat{x}$.

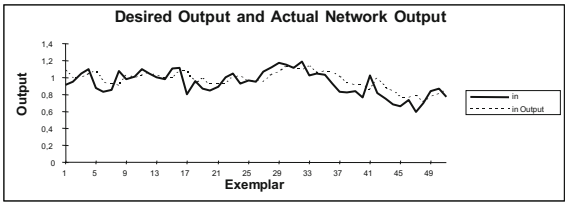

Fig. 6. Traffic intensity on a set of test points

6 Conclusions

The goal of the paper was to test the hypothesis that the use of the topological data analysis would make it possible to build traffic intensity prediction models due to finding additional characteristics that cannot be discovered by conventional data analysis. The data of network traffic intensity in a week's time were examined. The computations showed that the traffic intensity dynamics can be described by Betti numbers and Euler's characteristics. The algorithm using Euler's characteristics was used in the paper to build a model makes it possible to increase the prediction accuracy by an order of magnitude (as compared with methods not using Betti numbers). The paper gives the results of first steps towards the application of topological data analysis for predicting the network traffic intensity. The results proved the prospectiveness of further research in the field.

References

1. Heyman, D.P., Tabatabai, A., Lakshman, T.V.: Statistical analysis and simulation study of video teleconference traffic in ATM networks. IEEE Trans. Circuits Syst. Video Technol. **2**, 49–59 (1992)
2. Zhani, M.F., Elbiaze, H.: Analysis and prediction of real network traffic. J Netw **4**(9), 855–865 (2009)
3. Potapov, A.B.: Time-series analysis: when dynamical algorithms can be used. In: Proceedings of 5th International Specialist Workshop Nonlinear Dynamics of Electronic Systems, Moscow, 26–27 June 1997, pp. 388–393 (1997)
4. Edelsbunner, H., Letsscher, D., Zomorodian, A.: Topological persistence and simplification. Discret. Comput. Geom. **28**, 511–533 (2002)
5. Carlsson, G., Zomorodian, A.: Computing persistent homology. In: Proceedings of 20th Annual Symposium on Computational Geometry, pp. 347–356 (2004)
6. Haykin, S.: Neural Networks: A Comprehensive Foundation, 2nd edn. (2006)