Classification of Internet Traffic: A Distributional Data Approach

Sónia Dias[1,3][0000−0002−2100−2844], Paula Brito[2,3][0000−0002−2593−8818], and Paula Amaral[4][0000−0002−0563−3443]

[1] Instituto Politécnico de Viana do Castelo, Portugal sdias@estg.ipvc.pt
[2] Faculdade de Economia, Universidade do Porto mpbrito@fep.up.pt
[3] LIAAD-INESC TEC, Portugal
[4] NovaMath & Faculty of Science and Technology, Universidade Nova de Lisboa, Portugal paca@fct.unl.pt

Abstract. We address a classification problem where data are not single-valued, but distributions. The objective is to identify Internet traffic re-direction. Each observation consists of a block of 10 measurements of round-trip-times (RTT) measured at each of a set of *probes*, and is represented by the corresponding empirical distribution. The proposed approach relies on a method for discriminant analysis of distributional data that uses fractional programming, and where distributions are represented by quantile functions, under specific assumptions. A linear discriminant function is defined, that allows obtaining a score for each unit, in the form of a quantile function. This is then used to classify the units in *a priori* groups, using the Mallows distance. Results show that proposed approach works well, allowing for the identification of the diverted traffic.

Keywords: Classification · Histogram data · Multivariate statistics · Symbolic Data Analysis.

1 Introduction

Complex data, where "observations" are not single numerical values or categories, but include intrinsic variability, occur frequently. This is very pertinent in Data Mining applications where vast sets of data are collected, but data should be analysed at a higher level - take, e.g., the case of telecommunication companies, which record data on each call made (duration, etc.), but where the focus does not lie on individual calls but rather on client behaviour, and therefore information about the calls of each client, or specific group of clients, must be somehow aggregated. However, for each recorded feature, the observed variability inherent to each client or group should be taken into account; if one relies, as it is usually done, on central measures - means, medians, or modes - relevant information is irremediably lost. Symbolic Data Analysis (see, e.g. [1–3]) provides a framework allowing to represent and analyse data including inherent variability, and that go beyond traditional data models where one single value is recorded for

each unit on each variable. This has lead to the introduction of new variable types, where the observed "values" for each unit are not just single numerical values or categories, but finite sets of values, intervals or, more generally, distributions over a given domain. In this latter case, we say that we are in presence of distributional-valued variables.

In this work we are concerned with a problem of Internet traffic analysis, where each "observation" consists of multiple measurements, with the objective of identifying traffic diversion. Data are aggregated in the form of empirical distributions, for each measurement variable. We apply a method for classification of distributional data, that relies on the representation of distributions by the respective quantile functions, under specific assumptions, and uses a distance-based rule for class assignment [4].

The remainder of the paper is organised as follows: Section 2 describes the Internet traffic problem. In Section 3, we introduce histogram-valued variables and their representation, and define linear combination for these variable type. Section 4 details the linear discriminant method for histogram-valued data, which is then applied to the Internet data in Section 5. Section 6 summarizes the paper, pointing out avenues for future research.

2 The Internet Traffic Problem

We address a problem of identifying Internet traffic re-directions ("attacks"). [5]. For this purpose we use measurements obtained from a worldwide distributed probing platform, designed to detect routing variations based on round-trip-times (RTT) deviations inferred from disperse locations designed as *targets*. There are 12 *probes*, where the RTT are measured, and four *targets*, namely Chicago, Frankfurt, Hong-Kong, and London, here we focus on the London target. Regular traffic goes from *probe* to *target* and comes back to the *probe*; when an "attack" occurs, traffic is diverted through a *relay* before returning to the *probe*, and the RTT is typically larger. Table 1 lists the *probes*, *targets*, and *relays*. The setup is described in [5] and has been originally proposed in [10].

Table 1. Internet data: *Probes*, *targets* and *relays*.

Probes

Amsterdam	Chicago
Frankfurt	Los Angeles 2
Iceland	São Paulo
Milan	Viña del Mar
Sweden	Johannesburg 1
Israel	Johannesburg 2

Targets

Chicago	Frankfurt
Hong Kong	London

Relays

Los Angeles 1	Madrid
Moscow	São Paulo 1

The objective is to determine if a *target* is under attack. For each *probe*, an observation consists of 10 measurements of RTT. Summarizing these measurements by a central value, e.g. the mean, would lead to a high loss of potentially relevant information. We choose to represent each set of measurements by the corresponding empirical distribution, and we characterise each such distribution by a histogram defined on the two intervals [minimum , median[and [median , maximum]. Table 2 shows some units for the London *target*, one corresponding to a regular case, and the others to "attacks" through two different *relays*.

Table 2. Distributional data for London (partial view).

	Probe: Amsterdam	Probe: Chicago	...	Group
s_1	$\{[8.9, 9.2[, 0.5; [9.2, 9.4], 0.5\}$	$\{[88.0, 89.8[, 0.5; [89.8, 93.2], 0.5\}$	...	Regular
...	...	...	...	...
s_{9000}	$\{[143.0, 149.2[, 0.5; [149.2, 155.3], 0.5\}$	$\{[144.0, 148.65[, 0.5; [148.65, 153.4], 0.5\}$	...	Relay LA1
...	...	...	...	...
s_{9700}	$\{[38.5, 38.65[, 0.5; [38.65, 38.8], 0.5\}$	$\{[112.6, 116.25[, 0.5; [116.25, 120.3], 0.5\}$	...	Relay Madrid
...	...	...	...	...

Table 3. Number of units corresponding to regular and diverted traffic, by *relay*.

Target	Regular	Relay				Total
		LA1	Madrid	Moscow	São Paulo	Attacks
London	8569	681	567	782	835	2865

Table 3 presents the number of regular and diverted observations. To obtain a discriminant score for each unit, that allows for its classification as regular or attack, we define a discriminant function as a linear combination of the histogram-valued variables. The next section addresses this problem.

3 Histogram-valued Data

When the underlying domain of a distributional-valued variable is a compact subset of $\mathbb{R}$, we have a histogram-valued variable - see also [7]. In our problem, data take the form of empirical distributions of the RTT, and are represented by histograms.

Histogram-valued variables are formally defined as follows.

Definition 1. *Y is a histogram-valued variable when to each unit $i = 1, \ldots, n$ corresponds a histogram $Y(i)$ defined by a finite number of contiguous and non-overlapping intervals, each of which is associated with a (non-negative) weight. Then, $Y(i)$ can be represented by a histogram:*

$$H_{Y(i)} = \left\{ \left[\underline{I}_{Y(i)_1}, \overline{I}_{Y(i)_1}\right[, p_{i1}; \left[\underline{I}_{Y(i)_2}, \overline{I}_{Y(i)_2}\right[, p_{i2}; \ldots; \left[\underline{I}_{Y(i)m_i}, \overline{I}_{Y(i)m_i}\right], p_{im_i} \right\} \tag{1}$$

$p_{i\ell}$ is the probability or frequency associated with the subinterval $\left[\underline{I}_{Y(i)_\ell}, \overline{I}_{Y(i)_\ell}\right[$, m_i is the number of subintervals for unit i; $\sum_{\ell=1}^{m_i} p_{i\ell} = 1$; $\underline{I}_{Y(i)_\ell} \leq \overline{I}_{Y(i)_\ell}$ and $\overline{I}_{Y(i)_{\ell-1}} \leq \underline{I}_{Y(i)_\ell}$, $\ell = 1, \ldots, m_i$.

Each subinterval $I_{Y(i)_\ell}$ may be represented by its lower and upper bounds $\underline{I}_{Y(i)_\ell}$ and $\overline{I}_{Y(i)_\ell}$, or by its centre $c_{Y(i)_\ell} = \frac{\overline{I}_{Y(i)_\ell} + \underline{I}_{Y(i)_\ell}}{2}$ and half-range $r_{Y(i)_\ell} = \frac{\overline{I}_{Y(i)_\ell} - \underline{I}_{Y(i)_\ell}}{2}$.

$Y(i)$ may, alternatively, be represented by the inverse cumulative distribution function, the quantile function $\Psi_{Y(i)}$, under specific assumptions. Assuming that within each subinterval $\left[\underline{I}_{Y(i)_\ell}, \overline{I}_{Y(i)_\ell}\right[$ the values for the variable Y, for unit i, are uniformly distributed, the quantile function is piecewise linear and is given by

$$\Psi_{Y(i)}(t) = \begin{cases} c_{Y(i)_1} + \left(\frac{2t}{w_{i1}} - 1\right) r_{Y(i)_1} & \text{if } 0 \leq t < w_{i1} \\ c_{Y(i)_2} + \left(\frac{2(t-w_{i1})}{w_{i2}-w_{i1}} - 1\right) r_{Y(i)_2} & \text{if } w_{i1} \leq t < w_{i2} \\ \vdots & \\ c_{Y(i)_{m_i}} + \left(\frac{2(t-w_{i(m_i-1)})}{1-w_{i(m_i-1)}} - 1\right) r_{Y(i)_{im_i}} & \text{if } w_{i(m_i-1)} \leq t \leq 1 \end{cases} \tag{2}$$

where $w_{i\ell} = \sum_{h=1}^{\ell} p_{ih}, \ell = 1, \ldots, m_i$.

On the basis of quantile function representation, histogram-valued observations $Y(i), Y(i')$ may be compared by the Mallows distance (also known as $L2$-Wasserstein distance):

$$D_M(\Psi_{Y(i)}, \Psi_{Y(i')}) = \sqrt{\int_0^1 (\Psi_{Y(i)}(t) - \Psi_{Y(i')}(t))^2 dt} \tag{3}$$

Under the uniformity hypothesis, and considering a fixed weight decomposition (m_i constant, same weights, different intervals, as it is the case in our representation of the Internet data) we have [8]:

$$D_M^2(\Psi_{Y(i)}, \Psi_{Y(i')}) = \sum_{\ell=1}^{m} p_\ell \left[(c_{Y(i)} - c_{Y(i')})^2 + \frac{1}{3}(r_{Y(i)} - r_{Y(i')})^2 \right] \tag{4}$$

The Mallows *barycentric histogram* $\overline{\Psi_X}$ is the solution of the minimization problem $\overline{\Psi_Y} := \arg\min \sum_{i=1}^{n} D_M^2(\Psi_{Y(i)}(t), \Psi_{Y_b}(t))$. It is defined by the quantile function where the centres and half ranges of each subinterval ℓ are the classical mean of the corresponding centres $\overline{c}$ and half ranges $\overline{r}$ of all observations.

Covariance between two histogram-valued variables X and Y, based on the Mallows distance, has been defined in [9] as

$$cov(X,Y) = \frac{1}{n} \sum_{i=1}^{n} \int_0^1 \left(\Psi_{X(i)}(t) - \overline{\Psi_X}(t)\right)\left(\Psi_{Y(i)}(t) - \overline{\Psi_Y}(t)\right) dt \tag{5}$$

Assuming the Uniform distribution across the subintervals, this is written as

$$cov(X,Y) = \sum_{i=1}^{n}\sum_{\ell=1}^{m} p_\ell \left[(c_{X(i)_\ell} - \overline{c}_{X_\ell})(c_{Y(i)_\ell} - \overline{c}_{Y_\ell}) + \frac{1}{3}(r_{X(i)_\ell} - \overline{r}_{X_\ell})(r_{Y(i)_\ell} - \overline{r}_{Y_\ell})\right] \tag{6}$$

3.1 Linear Combination of Histogram-valued Variables

The space of quantile functions is a semi-vectorial space: the sum of quantile functions is still a quantile function, but the product of a quantile function by a scalar is quantile function only if this scalar is non-negative - otherwise a decreasing function is obtained, which cannot be a quantile function. For this reason, the linear combination of histogram-valued variables, represented by quantile functions, requires a special definition. In [6] a definition for linear combination of quantile functions is proposed that solves the problem of the semi-linearity of the space of the quantile functions. This is based on using both the quantile functions representing the histograms of the observed distributions and the quantile functions that represent the respective symmetric histograms - two terms per independent variable. Therefore, the non-negativity restrictions on the parameters do not imply a direct linear relation, while non-colinearity is ensured. A linear combination of histogram-valued variables $X_j, j = 1, \ldots, p$, is defined as (see [6]):

$$\Psi_Z(t) = \sum_{j=1}^{p} a_j \Psi_{Y_j}(t) - \sum_{j=1}^{p} b_j \Psi_{Y_j}(1-t), \text{ with } t \in [0,1]\,; a_j, b_j \geq 0 \tag{7}$$

4 Linear Discriminant Analysis of Histogram-valued Data

The linear discriminant function $S(i)$ that allows obtaining a discriminant score for each unit is hence defined by a linear combination of the observed histogram-valued variables as in (7):

$$\Psi_{S(i)}(t) = \sum_{j=1}^{p} a_j \Psi_{Y_j(i)}(t) - \sum_{j=1}^{p} b_j \Psi_{Y_j(i)}(1-t) \text{ with } t \in [0,1]\,; a_j, b_j \geq 0 \tag{8}$$

In the presence of s groups, the sum of the squared Mallows distance between $\Psi_{S(i)}(t)$ and the corresponding barycenter $\overline{\Psi_S(t)}$ may be decomposed in the sum of the squares and cross-products (SSCP) between groups and within groups, i.e,

$$\sum_{i=1}^{n} D_M^2(\Psi_{S(i)}(t), \overline{\Psi_S}(t)) = \sum_{k=1}^{s} n_k D_M^2(\overline{\Psi_S}(t), \overline{\Psi_{S_k}}(t)) + \sum_{k=1}^{s} \sum_{i \in G_k} D_M^2(\Psi_{S(i)}(t), \overline{\Psi_{S_k}}(t)) \quad (9)$$

with n_k the cardinal of group k, and the quantile functions $\Psi_{S(i)}(t)$ - *score*, $\overline{\Psi_S}(t)$ - *barycentric score* and $\overline{\Psi_{S_k}}(t)$ - *barycentric group score*. In matricial form, $\gamma' T \gamma = \gamma' B \gamma + \gamma' W \gamma$, where T is the matrix of the total SSCP, B and W are the matrices of the SSCP between-groups and within-groups, respectively.

As in classical linear discriminant analysis, the optimal parameter vector γ^* is estimated as

$$\gamma^* = \arg\max_{\gamma} \lambda = \frac{\gamma' \mathbf{B} \gamma}{\gamma' \mathbf{W} \gamma} \quad \text{subject to } \gamma \geq 0 \quad (10)$$

The obtention of the vector of parameters γ^* requires the optimization of constrained rational quadratic functions. This is a non-convex, hard optimization problem, for which it is easy to find a good solution but difficult to prove optimality. Optimisation is done by a Branch and Bound technique relying on Conic Optimization - see [4].

For more than two groups, successive discriminant linear functions must be derived, maximising λ under the additional constraints that they should be non-correlated with the previous ones (null covariance, see formula (6)).

Classification in the *a priori* groups is done using the Mallows distance: a unit s_i is assigned to the group for which the Mallows distance between its score and the score of the group's barycentric histogram is minimum.

5 Classification of the Internet Traffic Data

We first consider the two-class problem, consisting in separating the regular traffic from the "attacks".

Applying the method presented above, we obtained the linear discriminant function

$$\Psi_{S(i)}(t) = 0.29885\Psi_{Y_2(i)}(t) + 0.0085948\Psi_{Y_3(i)}(t) - 0.73097\Psi_{Y_4(i)}(1-t) + 0.0007634\Psi_{Y_5(i)}(t) - 0.00065739\Psi_{Y_6(i)}(1-t) - 0.12929\Psi_{Y_7(i)}(1-t) - 0.0019676\Psi_{Y_8(i)}(1-t) + 0.025075\Psi_{Y_9(i)}(t) + 0.95427\Psi_{Y_{10}(i)}(t) - 0.22277\Psi_{Y_{11}(i)}(1-t) - 0.00014969\Psi_{Y_{12}(i)}(1-t), \text{ with } t \in [0,1].$$

This shows that the *probes* that influence more the discriminant score (a quantile function) are (in this order) Y_{10} - Johannesburg1 , Y_4 - Frankfurt (in the opposite direction), Y_2 - Chicago, and Y_{11} - Johannesburg2.

The classification of each unit in one of the two groups is then done on the basis of the distance between the corresponding score and the scores of the barycenter of each of the two groups.

This lead to the results displayed in Table 4. We note that the proposed method successfully distinguishes the two groups, identifying more than 99% of the diverted traffic.

Table 4. Classification results for the two-class problem.

Target	Accuracy	Precision	Recall
London	0.9987	0.9983	0.9965

We further addressed the five-class problem, where we seek to identify "attacks" by *relay* from the regular observations. The hit-rates obtained on the basis of either four or just the first discriminant functions are displayed in Table 5. We observe that all four *relays* are well distinguished between themselves, and from the regular observations, with error-rates never above 0.053 when four discriminant functions are considered, but reducing to less than 0.0012 when using just the first one. In this particular problem, the results strongly suggest that the five classes are separable along just one (distributional) dimension.

Table 5. Internet data: hit-rates (%) for the five class problem.

Target	Nb. functions	Regular	LA1	Madrid	Moscow	São Paulo	**Global**
London	4	96.9891	97.9442	94.7090	96.1637	99.7608	97.0900
	1	100.0000	100.0000	100.0000	100.0000	99.8804	99.9999

6 Conclusion

We presented a classification method for numerical distributional data, where discriminant scores for each unit take the form of quantile functions. This score is obtained by an appropriate linear combination of the histogram-valued variables, where the model parameters are obtained by the optimization of a constrained fractional quadratic problem. This approach allowed identifying "attacks" in Internet traffic data.

Current research concerns the development of the method for the classification in more than two groups. For successive discriminant functions additional constraints must be considered. Furthermore, the relevance of the different discriminant functions must be assessed.

Acknowledgments. This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within projects UID/50014/2023 and UIDB/00297/2020.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bock, H.-H., Diday, E. (Eds.): Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data. Springer, Berlin (2000) https://doi.org/10.1007/978-3-642-57155-8
2. Billard, L., Diday, E.: Symbolic Data Analysis: Conceptual Statistics and Data Mining. John Wiley & Sons, Chichester (2006) https://doi.org/10.1002/9780470090183
3. Brito, P.: Symbolic data analysis: another look at the interaction of Data Mining and Statistics. WIREs DMKS, **4**(4), 281—295 (2014) https://doi.org/10.1002/widm.1133
4. Dias, S., Brito, P., & Amaral, P.: Discriminant analysis of distributional data via fractional programming. European Journal of Operational Research, **294**(1), 206–218 (2021). https://doi.org/10.1016/j.ejor.2021.01.025
5. Subtil, A., Oliveira, M. R., Valadas, R., Pacheco, A., & Salvador, P.: Detecting internet-scale traffic redirection attacks using latent class models. In: Proc. Tenth International Conference on Soft Computing and Pattern Recognition (SoCPaR 2018) vol. 10, pp. 370–380. Springer International Publishing (2020) https://doi.org/10.1007/978-3-030-17065-3-37
6. Dias, S., Brito, P.: Linear regression model with histogram-valued variables. Statistical Analysis and Data Mining, **8**(2), 75–113 (2015) https://doi.org/10.1002/sam.11260
7. Brito, P., Dias, S.: Analysis of Distributional Data. Chapman and Hall/CRC, Boca Raton, USA (2022) https://doi.org/10.1201/9781315370545
8. Irpino, A., Verde, R.: A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In: V. Batagelj, H.-H. Bock, A. Ferligoj, & A. Ziberna (Eds.), Data Science and Classification, Proc. IFCS'06 pp. 185—192. Springer, Berlin Heidelberg (2006)
9. Irpino, A., Verde, R.: Basic statistics for distributional symbolic variables: a new metric-based approach. Advances in Data Analysis and Classification, **9**, 143–175 (2015) https://doi.org/0.1007/s11634-014-0176-4
10. Salvador, P., Nogueira, A.: Customer-side detection of internet-scale traffic redirection. In: Proc. 16th International Telecommunications Network Strategy and Planning Symposium (Networks) pp. 1-5. IEEE (2014) https://doi.org/10.1109/NETWKS.2014.6958532