

# Classification of heavy-tailed data as differentiation service tool

**MARKOVICH, Natalia** *Institute of Control Sciences Russian Academy of Sciences, Russia,*  
markovic@ipu.rssi.ru

In order to improve the quality of service, one has to classify first the objects under the control.

Let us give the examples of a possible application of the classification procedure for the network traffic.

1. Mobile host service.

Reservation of capacity should be different regarding

a. applications (audio library, image browsing, video, etc);

b. the type of mobile host (the receiver, the sender, the receiver and the sender).

2. The http requests may be of different types:

a. Web pages (HTML); b. images; c. multimedia streams.

3. "Intelligent browser". is the classifier which can select what image it should load depending on the typical behavior of the user. More exactly, suppose the browser at first offers the user the information about the size of a picture. The user can ask the browser to show him a complete picture or to reject looking at this picture at all. Observing the work of the user during some fixed period of time one maintains two data sets: the sizes of rejected pictures (i.e., the ones which the user did not want to open) and the sizes of accepted pictures (opened by the user after the preliminary information of the browser). Then one can construct a classifier using the observations from two classes. Suppose, the separate observation of all sources (or situations) is available, for example, size files are measured. Then one can estimate the probability densities of the size files of the sources and provide a classifier. A classifier is a function that assigns the number of class to a value of a characteristic of the observed object. For example, one can separate the customers by the observation of the sizes of their http requests. Indeed, the classification has to be done with the minimal probability of a misclassification.

It is known, that the minimal risk of a misclassification is attained at the Bayesian classifier. The latter classifier assigns the object to some class if the multiplication of the probability density of this class of observations and the proportion of this class (a priori probability of the class) is maximal among all corresponded multiplications of other classes. Since the density is usually unknown, one has to use instead its estimate. Then the empirical Bayesian classifier is used. The risk of misclassification of the latter classifier is larger then the minimal risk of the Bayesian classifier.

Obviously, the problem is the density estimation.

The analysis of measurements of Web-traffic by statistical methods has shown that WWW characteristics (e.g., file sizes, durations of sub-sessions) are often heavy-tail distributed. It implies, that the "outliers" play a significant role in these data and cannot be excluded before the analysis like it is often recommended in robust methods which are stable with respect to contaminations of the data.

For finite and light-tailed distributions (i.e., those without heavy tails) a histogram is a good estimate of the corresponding density. But if the distribution is heavy-tailed, a histogram provides an absolutely misleading estimate in the "tail" domain. The same is true for most of the common non-parametric density estimates such as kernel, projection and spline estimates. In general, they have sharp peaks at "outliers" or over-smooth the density.

It is obvious that non-parametric (when the form of the distribution is not assumed) density estimates with good behavior at the tail domain are required to construct accurate classifiers. Since the object can arise in the tail domain as well as in the body, a tail estimator with good properties is principal for the classification.

To improve the estimation, it was proposed to transform the data (the simplest example gives the logarithmic transformation) before the application of a specific non-parametric method. For this purpose, appropriate parametric or non-parametric approximations of the distribution function may be used. Then one can estimate the density of the original data by the reverse transformation of the density estimate of the transformed data.

The special adapted transformation that uses the Generalized Pareto distribution as a reasonable approximation of the true distribution function and a triangular distribution as a target distribution is considered. This transformation provides the re-transformed density estimates that are stable to minor perturbations in the transformation and maintain better the tail decay rate of the true density.

A simulation study of some non-parametric density estimates to solve a classification problem in the case of different heavy-tailed distributions is presented. An analysis of real data generated by Web sessions is provided.