UDC 004.8

Serge Dolgikh, M.Sc., Research Associate

PARAMETER-LESS HISTOGRAM-SCALE METHOD OF BANDWIDH ESTIMATION IN DENSITY-BASED CLUSTERING

Density clustering methods such as MeanShift, DbScan [1] and their numerous extensions and variants are commonly used in the problems of information science in the analysis of complex distributions the behavior and parameters for which may not be known from the outset. These methods allow to identify the distribution of density in general datasets with minimal assumptions about the nature of the underlying distribution. However, the result of application of these method in the general case may depend significantly on the choice of the value of the bandwidth parameter *h*, which may not be straightforward for arbitrary data and require non-trivial additional analysis.

Whereas methods for estimation of bandwidth parameter in general datasets do exist including those based on empirical formulas [3], programmatic implementations, grid search and other methods, they have some well-known limitations. For example, some of the methods require certain non-trivial assumptions about the distribution correctness of which may not be known without additional analysis; others depend on the choice of a specific kernel function or additional parameters such as quantile value. Grid search over the range of possible values in the parameter space is a common approach to overcome these challenges however it can be expensive and time consuming. Hence, it would be of benefit to identify an approach to estimation of the bandwidth parameter in general data of arbitrary type and origin that would not require additional assumptions about the data.

Objective: To develop, implement and verify a method of estimation of the bandwidth parameter in density clustering that is applicable to data of general type and does not require additional assumptions about the distribution of the data and / or values of other parameters.

The approach proposed in this work is based on evaluation of the scale dependency of the histogram of the distribution for general data with values in a multi-dimensional space. A functional analysis of the histogram dependency on the scale allows to estimate important characteristics of the distribution such as characteristic scale of variation of density. An improvement to the existing methods is that histogram analysis can be performed only with the analyzed dataset and does not depend on any additional assumptions about the distribution type or parameters.

The method of histogram bandwidth estimation is based on calculation of the slope of the histogram-scale function $H(s,\ p)$ where the scale *s* denotes the size of a bin in the multi-dimensional histogram of the dataset, and $H(s)$, the total number of bins with density above average, $d > D_0$. A calculation of the histogram-scale function can be performed straightforwardly on a general dataset with existing methods that allows to estimate the derivative of the histogram-scale function and identify the characteristic points in the range of variation of scale where the slope undergoes significant change. Such a change indicates that the bins have achieved characteristic scale of the density variation in the distribution identified by a critical point, or a set of critical points $\{ S_k \}$.

The algorithm of the proposed method tracks the change in the slope of the histogram-scale function *H(s)* by detecting the set of critical points of discontinuity of the histogram-scale derivative *H'(s, p)* in the range of scale for the given dataset as:

$$S_k := \frac{\partial H(s)}{\partial s}(S_k + \varepsilon) > \frac{\partial H(s)}{\partial s}(S_k - \varepsilon)$$

The characteristic point(s) $\{ S_k \}$ of *H(s)* are therefore directly related to the characteristic scale of the density distribution of the data in the dataset and the bandwidth parameter *h* that can be approximated as αS_k, where $\alpha \subset]0, 1]$, a factor. Hence, as a result of application of the method with a general set of data of virtually arbitrary type and origin, it is possible to obtain a small set of suggested bandwidth values. The method was verified with data of different types such as Internet packets and images.

The proposed general method that does not make assumptions about the characteristics and parameters of the distribution of data in a general dataset and can be useful in the analysis of general data with unknown law and parameters of the distribution.

List of references:

1. Fukunaga K., Hostetler L.D. (1975): The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on Information Theory. 21(1), 32 – 40.
2. Park, B.U.; Marron, J.S. (1990). Comparison of data-driven bandwidth selectors. Journal of the American Statistical Association. 85 (409), 66–72.