DOI 10.18699/SBB-2022-659

Multidimensional analysis, dimension reduction, categorization with statistical approach – stability and reproducibility

Karpenko D.V.
National Research Center for Hematology, Moscow, Russia
d_@list.ru

Key words: multidimensional data analysis, binary clustering, secondary parameters, prognosis self check, statistical weight, statweight

Motivation: The analysis of modern medical and biological data arrays formulates various tasks in the multidimensional space of experimental data. Categorization tasks are associated with the formation of forecasts; dimensionality reduction can also solve the problem of extracting key parameters that describe groups. Modern powerful tools using neural networks involve the use of large training datasets. However, medical data are often characterized by their paucity and, in some cases, even blurring of the boundaries of the original groups and the presence of transitional elements. In most cases, we even don't know if groups differ by the given set of parameters. The complexity of the data obtained at different times in different sources adds more difficulties.
Aims: Create a unified, automated clustering and group description tool capable of processing mosaic data, which is resistant to single outliers. Additional tasks are formulated as taking into account individual fluctuations of measurements, taking into account the uncertainty of the group, and taking into account the pairwise relationships of parameters.
Methods and Algorithms: The boundaries of the dynamic range of data are determined, data outside this range are determined as fixed values corresponding to the maximum or minimum limit, and overestimated error values are signed to it. Values approaching the boundary of the dynamic range are retained, but are associated with a larger error. The data are converted to a logarithmic scale, and the group average and the scatter of values are calculated for each parameter separately. The mean and spread are calculated recursively using an exponential statistical weight (statweight) determined from the deviation from the mean and measurement error divided by the group scatter. After calculating the average for the group, the parameters of paired bonds are formed: in a two-dimensional space, a pair of parameters is determined – at what angle and with what spread the point of the group is observed relative to the center of the group. The set of obtained angles is interpreted as a set of independent parameters, thus the primary and secondary parameters become equal. Next, the problem of binary clustering is solved for each parameter separately through the question of which group the point differs more from. The difference is calculated as the difference between the value of the point and the group divided by the total error of the related objects, and the ratio of differences from the first and second groups is compressed in the range from 0 to 1. The calculated average for all parameters answers the question of binary clustering or forecasting. At this stage, a set of parameters that are optimal for group separation is selected. The stability of the prognostic model is checked by alternately excluding one element of the group and evaluating its prognostic value, and so on for all elements. Thus, the influence

of individual elements on small groups is checked. Early versions of algorithm was applied previously [1, 2].

Results: An algorithm was developed and tested in Python v.3.7 in Microsoft Windows. The program allows to process incoming data as separate blocks in Excel list, scripts perform the primary sorting of incoming data. The output is a list of parameters arranged according to their significance for the difference between the target groups and a graph in the form of dual ratios of points between the target groups. The mode of checking the stability of categorization through the exclusion of elements from groups results in a graph, where each point is marked separately and is located in the range from 0 to 1, depending on the differences from which group are more significant.

Conclusion: Dual clustering and space reduction are the development of the idea of scalar projection onto the axis connecting the centers of groups. A significant rethink is the comparison of distances from a point to a group in terms of the errors in determining the point and the group. Creating new parameters from the original ones is an attempt to define primitive relations between parameters and consider them as independent elements. It can be seen that the least squares method with a linear interpretation of dependence or the principal components analysis have similar foundations, namely, the selection of linear dependences. However, the schemes for introducing the used approaches with statweight, exclusion testing and binary prediction into these tools are not obvious. The scheme used made it possible to compare directly and equally the primary and secondary parameters. The creation of new parameters in the original multidimensional parameter space for small groups should be done with caution, since the data features, which are found that way, may be sporadic. The algorithm is able to find differences and reliably distinguish elements of groups, but when trying to carry out the proposed check, there is a strong inconsistency and the absence of a reliable predictive conclusion as a result of comparing insufficiently distinguishable groups. However, the algorithm and its checks perform well, when comparing sufficiently distinct groups. It is important to emphasize the significance of the proposed verification approach not only for the proposed algorithm, but also for other automated calculation schemes.

References

1. Petinati N., Kapranov N., Davydova Y., Bigildeev A., Pshenichnikova O., Karpenko D., Drize N., Kuzmina L., Parovichnikova E., Savchenko V. Immunophenotypic characteristics of multipotent mesenchymal stromal cells that affect the efficacy of their use in the prevention of acute graft vs host disease. *World J Stem Cells.* 2020;12:1377-1395. doi: 10.4252/wjsc.v12.i11.1377.
2. Petinati N.A., Bigildeev A.E., Karpenko D.S., Sats N. V, Kapranov N.M., Davydova Y.O. et al. Humoral Effect of a B-Cell Tumor on the Bone Marrow Multipotent Mesenchymal Stromal Cells. *Biochemistry. (Mosc).* 2021;(86):207-216. doi: 10.1134/S0006297921020097.