Contribution ID: **6** Type: **not specified**

Estimates of the learning set size for k-NN and IINC separating methods in the high energy physics.

Monday, 18 June 2018 15:30 (20 minutes)

Reliability of separation methods based on learning with the teacher (supervised learning) is primarily studied by verifying the independence of the achieved results on selected parts of data sets used. For this purpose, both data exploited in the process of separator parameters settings as well as independent test data are used. For example, the first one data are frequently used in so called cross-validation and the second one in the test of over-learning.

More sophisticated methods of verifying the reliability of learning methods with a teacher exploits additional knowledge of statistical characteristics of the data processed. These expected statistical characteristics are tested by standard statistical procedures (as an example can serve the statistical distribution of the mutual energy of $M_{b\bar{b}}$ pair in the decaying tree of $p\bar{p}$ collision).

All of these methods are based on the properties of the processed data only and do not give any assessment of the suitability of the method used to process the particular data. At the same time these methods do not provide any information regard to appropriate amount of separated data or convenient complexity of the separating model.

Concurrently an exact theory of PAC-learning has been developed for supervised learning methods. PAC-theory provides quantitative relationship between the number and dimension of the processed data on the one hand and the appropriate size of the parametric space of the separation methods on the other hand, under predefined conditions on expected quality and reliability of separation.

In our contribution we will show the application of the PAC-theory to separation methods of k-NN and IINC type. We will obtain necessary characteristics of these methods in the terms of PAC-learning and indicate application of derived data size estimates in the field of analyzing data produced by elementary particle detectors.

Primary author: HAKL, Frantisek (Institute of Computer Science CAS, Prague, Czech Republic)

Presenter: HAKL, Frantisek (Institute of Computer Science CAS, Prague, Czech Republic)

Session Classification: Data processing in HEP