Neural Networks as Tool for Event-by-Event Analysis in Astroparticle Physics

Ashot Chilingaryan, Ararat Vardanyan

Cosmic Ray Division, Yerevan Physics Institute, Armenia

Abstract:

Modern arrays of particle detectors covering a large area are measuring different parameters of numerous secondary products of the primary cosmic ray interactions with the atmosphere. Only a simultaneous measurement of a large number of independent parameters in each individual Extensive Air hower (EAS) can yield reliable information to reconstruct the Primary Cosmic Radiation (PCR) particle mass and its energy as well as the characteristics of strong interaction with atmosphere nuclei.

To make the conclusions about the investigated physical phenomenon reliable and significant, it is necessary to develop a unified framework of statistical inference, based on nonparametric models, in which various nonparametric methods (Bayesian decisions, Neural Networks models, Feature extraction, etc, . . .) would be incorporated.

In the paper our approach for coherent solution of data analysis problems encountered in Astroparticle Physics experiments is presented.

Introduction :

Currently there is no universally accepted theory of the predictive learning. Statistical learning theory, developed by V.Vapnik [1], based on theoretical analysis Empirical Risk Minimization (ERM) is a theory for nonparametric dependency estimation with finite data. The Vapnik-Cervonenkis (VC) theory drives necessary and sufficient conditions for consistency of generalization from finite set of examples. The

generalization ability of a learning algorithm depends both on the possibility to find the particular function describing the examples on on the measure of the complexity of the used family of approximated functions. Classical notions of complexity (number of free parameters or degrees of freedom) fail to account for applications to the functional families with infinite number of members like Neural Networks models. The VC theory, introducing the so-called VC dimension (capacity) as a measure of complexity, can deal and specify functional families ever with infinite number of members. Therefore the VC theory provides conceptual framework for the setting bounds for model complexity control. The issue of the model complexity control is of crucial importance for the practical application of learning algorithms. Of course, there is still much work needed to bridge the gap between theory and practical applications. However, such empirical approaches for selecting the optimal Network solution to avoid overfitting problem, as prediction risk estimation [3] and median networks committee [4] could be addressed in rigorously defined mathematical scope.

In the cosmic ray physics the main technique of physical analysis is the Monte Carlo Statistical Inference (MCSI), the detailed simulation of the CR traversal through the atmosphere and the experimental installation with a following comparison of the multivariate simulation and experimental data. Actually, an algorithm is constructed, which describes EAS development and registration of its different components on the observation level, which is based on a certain family of models of the physical processes investigated.

MCSI is a process (see picture 2.5). It takes requirements specification (basic physics, experimental techniques, data analysis techniques), it generates families of models to meet this specifications and it synthesizes a *priori* knowledge and experimental results to create new knowledge.

Complexity of the MCSI is determined by its multifunctionality, adaptability and flexibility - attributes that one best realized in Neural Network models.

Neural models captures the statistics of processes directly from data vectors - collection of "pseudoexperimental" variables, corresponding to all significant variations of the model input parameters.

Here in lies MCSI exibility. It allows the input vectors to be formed directly from initial measurements or from reconstructed EAS parameters.

Neural methods are universal and can deal with very big input vectors. A common complaint about nonparametric techniques is the dependence of the results on the purity and finiteness of training sets (small training samples effects). However, due to the inherent robust characteristics of Neural Network (generalization ability), results from neural analyses are relatively insensitive to modest impurities in the training sets.

MCSI incorporates and uses such advanced nonparametric methods as fuzzy Analysis, Nonparametric Boundary Analysis, Adaptive Multivariate Density Estimation, Fractal Dimensionality Analysis, etc, ... For net training the Evolutionary Algorithms are used, Stopping Rules, based on the Prediction Error estimation and Committee method provide high level of generalization, avoiding overstraining errors. For the training of very big networks hardware accelerators (neurichips) are used.

Statistical Learning Theory:

The overall scheme of learning from examples can be defined as following [2]:

1. Random event generator, drawn independently from a fixed but unknown distribution mixture;
2. A supervisor (absolute decision rule) that returns an output vector for every input vector, according to a unknown, but also fixed conditional distribution function;
3. A learning machine (algorithm) capable to implement a number (may be infinite) of different approximated functions.

The problem of learning is that of choosing the appropriate set of functions, and then particular member of this family, which predicts the supervisor's response in the best way (optimal decision rule). The selection is based on the training set (sample), of independent and identically distributed observations presented to the supervisor.

Usually, for experimental physics data analysis, the Likelihood Function cannot be written explicitly, and we deal with implicit, nonparametric models, for which no parametric form of underlying distribution is known, or can be assumed [14]. Nonparametric methods use much less stringent assumptions about population than those made in parametric statistics. Usually the underlying population distribution is assumed to be continuous only. Of course this assumption is rather mild comparing with the very specific assumptions made in parametric case.

Let us consider the stochastic mechanism (A, p) which generates the observation v in a multivariate feature space $-\nu, \mathrm{v}$ is a d-dimensional vector of EAS parameters measured experimentally. We assume that observations are randomized and can be described by some conditional probability density function depending on the primary particle type. The feature space ν covers possible acceptable values of EAS parameters including cuts on shower age and size, etc. The basic state space A consists of different primary nucleus. The appropriate statistical model describing EAS initiated by various primaries is the probability mixture model:

$$p(\nu) = \sum_{k=1}^{L} P_k \, p(\nu / \mathrm{A}_k) \ . \qquad (1)$$

The proportions (frequencies) of the probability mixture P_k of events in each category, A_k, determine the mass composition of the primary flux. Unfortunately, we don't know the full statistical description (conditional probability density functions $p(\nu / A_k)$ of how nature produces EAS from incident primaries, that is why, to determine the mutual probability measure on the direct product of A and ν spaces the total Monte-Carlo simulation of the EAS development in the atmosphere and in detectors is performed, including experimental data registration and reconstruction of EAS parameters for different primaries and alternative strong interaction models in a wide energy range. The problem is how to introduce the probability measure in the primary particle parameters space τ (K-dimensional metric space). Usually following parameters are used as input for Monte-Carlo simulation program:

- Primary type;
- Primary energy;

- Angles of incident; Strong interaction model (one of the CORSIKA alternatives [15]).

Of course, we've to implement the physical restriction and define the bounded subspace of τ, from which we randomly take the mesh points $(t_i, i=1, M)$, M is number of simulation trials. The primary particle classes will be restricted by 5 groups, including all primaries from proton to iron. The set of corresponding dimensional $(u_i, i=1, M)$vectors obtained in simulations is an analog of the experimentally measured values of $(v_j, i=1, M_{exp})$, where M_{exp}, is number of detected events. But, as opposed to experimental data, it is exactly known which primary particle was used in simulations. These, *labeled* events include a *priori* information about dynamics of the EAS development and registration with inherent uctuations. All statistical variability of events belonging to the definite class is expressed in a nonparametric form, in form of simulation trials. The sequence $(u_i, t_j), i=1, M_j$; j = 1, L,t- is the class index, is generated by CORSIKA simulation program [15] and consists of L classes each contaning M_j Simulation trials.This "controlled" stochastic mechanism we denote by $(A, \tilde{p})$ and will refer to it as training sample (TS). The training, sample is the basis of all statistical procedures in applied Bayesian and neural approaches. Usually we denote a TS by A k or explicitly by the primary group - P, O, . . . ,Fe.

The corresponding distribution mixture model takes the form:

$$\tilde{p}(v) = \sum_{k=1}^{L} \hat{P}_k \hat{p}(v / A_k) \qquad (2)$$

Of course this substitution of unknown conditional density function $p/(v/A_k)$ by "simulation" analog $\tilde{p}/(v/A_k)$, estimated by means of training sample $\{u_i, t_i\}$, is only valid if used model is adequate. And validation of the model remain the most crucial and yet unsolved problem for EAS data analysis. For reliable estimation of conditional densities we'll need significant amount of training trials to cover all intrinsic variations of measurable EAS parameters and completely represent all categories (primary nucleus). Since both physical processes of particle production and those of registration are stochastic, only by careful measurement of probabilities we can gain an understanding of the EAS phenomena. We can't expect simple solutions, as multidimensional distributions of EAS parameters overlap significantly and any decision on primary particle type and it's energy will contain uncertainty. The only thing we can require when classifying a distribution mixture is to minimize the losses due to incorrect classification to some degree and to ensure use of a priori information completely. Such a procedure is the *Bayes decision rule with nonparametric estimation of the multivariate probability density function*

Bayesian Decision Rules :

The Nonparametric Bayesian decision rule have a form [16]

$$\tilde{A} = \eta(v, A, \tilde{p}) = \arg\max_i \{C_i \hat{p}(A_i / v)\}, i = 1,..., L. \quad (3)$$

where C_i are the losses connected with $\tilde{A}$ decision, $\tilde{p}(A_i, v)$ is the nonparametric estimate of the *a posteriori* density, connected with conditional ones by the Bays theorem:

$$\hat{p}(A_i / v) = \frac{\hat{P}_i \hat{p}(v / A_i)}{\hat{p}(v)}. \quad (4)$$

Finally, substituting the a posteriori densities by the conditional ones we get the Bayesian decision rule in the form

$$\tilde{A} = \arg\max_i \{C_i P_i \hat{p}(v / A_i)\}, i = 1,..., L. \quad (5)$$

Provision is made to avoid statistical decision if all classes are very far from experimental events (outliers problem). If

$$\hat{p}(v / A_i) < ST \quad for\ all\ i = 1,..., K, \quad (6)$$

then the "outliers message" is send to output stream. ST is, so called, Strangeness Criteria, usually set to a small number.

The Nonparametric Likelihood Ratio for classes A 1 ; A 2 and experimental event v can be represented as:

$$LR(v) = \frac{\hat{p}(v / A_1)}{\hat{p}(v / A_2)}. \quad (7)$$

The nonparametric Log-likelihood function for $k-th$ class takes the form:

$$\ell_k = \sum_{i=1}^{M} \ln \hat{p}(v_i / A_k), k = 1, L, \quad (8)$$

where M is number of experimental events. The negative of Log Likelihood function is also calculated; the smaller values will correspond to most probable model.

Nonparametric Probability Density Estimators :

To estimate conditional densities, we use Parzen kernel [17, 18] and K Nearest Neighbors (KNN) methods [19, 20] with automatic adaptation of the method parameter (kernel width - for Parzen estimate, and number of neighbors - for KNN estimate)[21]. Several probability density estimates corresponding to different values of parameters are calculated imultaneously. Then the obtain sequence is ordered and the median of this sequence is chosen as final estimate. Depending on the intrinsic probability density in the vicinity of point v, where the density is estimated, due to stabilizing properties of the median, each time the best estimate will be chosen [22]. The Parzen kernel probability density is estimated by:

$$\hat{p}(v/A_i)=\frac{|\sum_i|^{-0.5}}{(2\pi)^{d/2}h^d}\sum_{j=1}^{M_i}e^{-r_j^2/2h^2}\omega_j, i=1,\ldots,L, \sum_{j=1}^{M_i}\omega_j=1 \quad (9)$$

where d is the feature space dimensionality, M_i is the number of events in the $i-th$ TS, r_j is the distance from exprimental event V to the $j-th$ event of the TS in the Mahalanobis metric

$$r_j^2=(v-u_j)'\sum_i^{-1}(v-u_j), \quad (10)$$

where $\sum_i$ is the sampling covariance matrix of the class to which u_j belongs, ω_j are the event weights, h is the kernel width (parameter controlling the degree of the "smoothness" of an estimate). The K nearest neighbors estimate takes the form:

$$\hat{p}(v/A_i)=\frac{k-1}{M_iV_k(v)}. \quad (11)$$

where $V_k(v)$ volume of a d-dimensional hypersphere containing the k nearest neighbors to the experimental event v,

$$V_k(v)=V_d|\sum_i|^{1/2}r_k^d, V_d=\frac{\pi^{d/2}}{\Gamma(d/2+1)}, \quad (12)$$

where r_k is the distance to the $k-th$ nearest neighbor of v, $\Gamma(.)$ is the gamma function. $|\sum_i|$ is the determinant of the covariance matrix of the class to which the $k-th$ neighbor belongs.

Bayes Error Estimation:

The classification methods, like all the statistical ones, include a procedure quality test as a necessary element. The most natural measure for quality test is the error probability which depends on both the degree of overlapping of alternative multivariate distributions and the decision rule being used:

$$R^B=E\{\theta[\eta(v,A,p)]\}=\int v\,p(v)\,dv, \quad (13)$$

where

$$\theta[\eta(v,A,p)]=\begin{cases}0, \text{for correct classification},\\ 1, \text{otherwise}\end{cases} \quad (14)$$

The mathematical expectation is taken over the whole *d*-dimentional feature space *V*. In other words the Bayes error is a measure of the overlapping of alternative distributions in the feature space *V*, *e.g.* the expected proportion of the "incorrect" classification. Since we do not know to which class experimental vectors belong, we obtain an estimate of R B via the TS:

$$\hat{R}^B=E\left\{\frac{1}{M_{TS}}\sum_{i=1}^{M_{TS}}\theta\{\eta(u_i,A,\tilde{p})\}\right\}, \quad (15)$$

i.e. we classify the $\{u_i\}, i=1, M_{TS}$ and check the correctness of the lassi_cation over the index of the class $t_j, j=1, L$ The expectation is taken over all ossible

samples of volume M_{TS}. However, as numerous investigations have shown (e.g.,[23]), this estimate is systematically biased and hence, a one-leave-out-for-a-time estimate is preferable:

$$\hat{R}^{e} = \frac{1}{M_{TS}} \sum_{i=1}^{M_{TS}} \theta\left\{\eta\left(u_i, A, \tilde{p}_{(i)}\right)\right\}. \tag{16}$$

where $\left(A, \tilde{p}_{(i)}\right)$ is a TS with a removed $i\text{-}th$ element, which is classified and then "returned" to the sample. This estimate is unbiased and has an ssentially smaller m.s. deviation compared with other estimators[24]. The advantage of $\tilde{R}^{e}$ is especially notable when the feature space has a high dimensionality. Note, that we have the possibility to estimate the error probability of various types by classifying various TS classes $-\{u_i, t_j\}, j = 1, L.$ By R_{ij}^{e} (or simply R_{ij}) we denote the probability of classifying the $j-th$ class events as belonging to the $j-th$ class (misclassification). By R_{ii} the "true" classification robability will be denoted. For EAS classification according to 5 primary groups, each
element of the "classification matrix" have to be determined, using the Bayes risk estimate (16).

$$\begin{pmatrix} R_p \to p & R_p \to \alpha & R_p \to 0 & R_p \to si & R_p \to fe \\ R_\alpha \to p & R_\alpha \to \alpha & R_\alpha \to 0 & R_\alpha \to si & R_\alpha \to fe \\ R_0 \to p & R_0 \to \alpha & R_0 \to 0 & R_0 \to si & R_0 \to fe \\ R_{si} \to p & R_{si} \to \alpha & R_{si} \to 0 & R_{si} \to si & R_{si} \to fe \\ R_{fe} \to p & R_{fe} \to \alpha & R_{fe} \to 0 & R_{fe} \to si & R_{fe} \to fe \end{pmatrix}$$

This matrix presents accumulate a priori knowledge on the possibility of data classification into 5 categories. The overall index reecting the goodness" of features used is following index of separability

$$(17)_G = \left(\prod_{i=1}^{L} R_{ii}\right)^{1/L}. \tag{17}$$

This averaged product of diagonal elements represents the "mean" robability of true classification into L categories. The separability index, of course, is directly connected with Bayes error.

Feed-Forward Neural Networks:

Feed-Forward Neural Networks (FFNN) represent very simple structures composed of processing elements (nodes) and connections (weights). FFNN belongs to the general class of non-parametric methods that do not require any assumption about the parametric form of the statistical model they use. The central issue of FFNN is implementation of the bounded mapping [25]:

$$f: U \subset R^{n1} \to R^{n2}, \tag{18}$$

from a bounded subset V of n1 dimensional Euclidean space to a bounded subspace f[V] of n2 dimensional Eucledian space (usually n1>n2). The special case of such

mapping when n1=1, constitutes the classification problem. Of course, for real live problems it is impossible to define non- overlapping devision of V corresponding to different categories, but using the examples of mapping action, a Network configuration can be urned to minimize the misclassification errors nesr to minimal achievable Bayes error (13).

The net architecture considts of L layers each having K nodes. The first layer consists of N1 elements that simply accept the components of input vector v and distribute them, without modification, to the all of the nodes of the secound layer. The nodes of the secound layer calculate a weighted sum of all inputes and then transform it to the third layer, and so on till the output layer with N2 nodes is reached. The output of a FFNN different classes of TS from each other as much as possible.

Therefore the "goal" output $O^{goal}(k)$ for events of the $k-th$ category could be chosen as follows:

$$O_k^{goal} = \frac{k-1}{K-1},\ k = 1, K. \tag{19}$$

where K is total number of classes. For the multi-way classification one can define a set of non overlapping bounded intervals in (0-1) for each category. This sequence of bounded non-overlapping sets $O_k, k = 1, K$, along with the chosen "goal" values (located within corresponding subsets), will determine the mapping into the K class labels:

$$O(u) \subset O_k \rightarrow u \ \ belongs\ to\ k_{th}\ category. \tag{20}$$

The objective (error) function to be minimized is simply the discrepancy of apparent and target outputs over all training samples (so called classification score):

$$Q = \sum_{k=1}^{K} \sum_{j=1}^{M_k} \omega_k (O_k^j - O_k^{goal})^2, \sum_{k=1}^{K} \omega_k = 1. \tag{21}$$

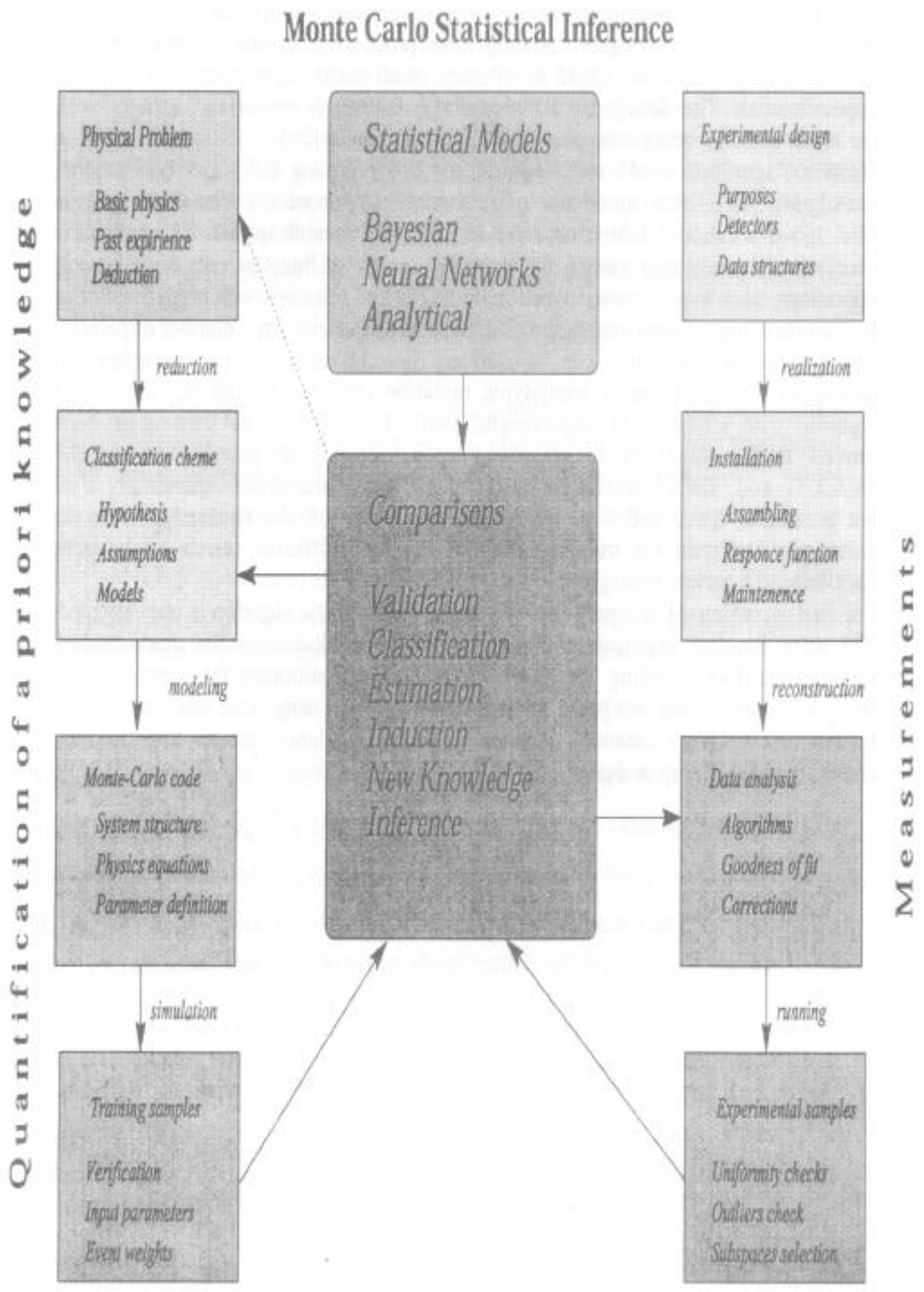
where O_k^j is the actual output value for the $j-th$ training event, belonging to the $k-th$ class and the O_k^{goal} is the target value for the $k-th$ class output, where K is number of categories and M_k is the number of examples for the $k-th$ class.

The ω_k weight cofficients controls the "contribution" of each particular class of TS to the overall error function. For the identification of the primary type by EAS observable, usually intermediate nucleus (oxygen class) with masses between the lightest (proton class) and heaviest with significant abundance (iron class) are trained worse compared with edge classes.There are two possibilities of checking the classification accuracy of middle categories. First of all we can enlarge the category *acceptance* region O_{middle}, (*a posteriori* solution)(20). And, second, the corresponding weight value in error function could be enhanced before starting net training (*a priory* solution) (21).

Net Training:

The only information to "train" network for "nonliner" mapping is contained in a priory given paires $-(t_i, u_i), i = 1, M$, where M is the number of training events.

During the minimization procedure the calculated differences between the actual network output and the desired output are used to adjust the weights.
The back-propagation (BP) algorithm of neural network training is one of the most important historical developments in neurocomputing. The simple rule (based on gradient descent) of wieghts updating after proccesing of one or more training examples in principle will lead to arbitary small mean square error of function approximation. The family of BP algoritms is realized in numerous packages, with the Jetnet package being most popular in HEP community[26].
Generic Algorithms (GA) and Evolutionary Programming (EP) are boath search technigues based on an simulation of the evolutionary processes. The challenge is to find "good solutions" (chromosomes) in very large search spaces. GA employ the successive reproduction among an assembly (pool) of fittest parents using generic operations such as crossover, inversion, mutation and selection with predefined rules for constructing of next generation. Different m:n scenarios (m – number of parents, n – number of offsprings) can be realized.The current best chromosome (parent) undergoes the zero mean phenotypic mutation (realized by the random search algoritm with return at an unsuccessful step). This kind of net training has been proved to be much more effective than simple random search algoritms.The MULLTI and SINGLE modes are designed for random search correspondingly in all net parameter space and – to make random change of also randomly chosen net parameter. Different net training scenarios combine different search modes with various search parameters.
For fast scanning of the net weights space a deterministic algoritm is implemented. The error function is calculated in each point of the multidimensional quasi-random sieve[27] uniformly filling the N-dimentional cube. Positioning the sieve center at the previously found best point, and subsequently decreasing sieve size, we'll arrive to the best net.Very essential question of scale invariance can be addressed by changing value of step in described above SOBOL mode.


Monte Carlo Statistical Inference
Quantification of a priori knowledge
Measurements
Physical Problem
Basic physics
Past expirience
Deduction
reduction
Classification cheme
Hypothesis
Assumptions
Models
modeling
Monte-Carlo code
System structure
Physics equations
Parameter definition
simulation
Training samples
Verification
Input parameters
Event weights
Statistical Models
Bayesian
Neural Networks
Analytical
Comparisons
Validation
Classification
Estimation
Induction
New Knowledge
Inference
Experimental design
Purposes
Detectors
Data structures
realization
Installation
Assambling
Responce function
Maintenence
reconstruction
Data analysis
Algorithms
Goodness of fit
Corrections
running
Experimental samples
Uniformity checks
Outliers check
Subspaces selection

References:

1. Vapnik, V.N., *Estimation of Dependencies based on Empiric data*, Moscow, Russia: Nauka, (1979), 448 pp, (in Russian); English translation, New York: Springer-verlag, (1982), 400 PP.
2. Vapnik, V.N., IEEE Trans. On NN, 10, (1999), 988.
3. Chilingarian, A.A., Vardanian, A.A., This Proceeding.
4. Chilingarian, A.A., Vardanian, A.A., This Proceeding.
5. Chilingarian, A.A., Zazyan, H.Z., Yad.Fiz, 54, (1991), 128.
6. *ANI collaboration*, NIM, A323, (1992), 104.
7. Chilingarian, A.A., Computer Physics Communications, 54, (1989), 381.
8. Chilingarian, A.A., Zazyan, H.Z., IL Nuvo Cimento, 14C, (1991), 555.
9. Kampret, K-H., Proc. 26 ICRC (salt Lake City, 1999), 3, 159.
10. Vardanian, A.A., et.al, Proc. Workshop ANI 99, (Nor-Amberd, 1999).
11. Chilingarian, A.A., et.al, Proc. 26 ICRC (Salt Lake City, 1999), 1, 226.
12. Kalmykov, N.N., Ostapchenko, S.S., Pavlov, A.I., Nucl.Phys.B, 52B, (1997), 17.
13. Chilingarian, A.A., Analusis and Nonparametric Inference in High Energy Physics and Astrophisycs Experiments, Refrences Manual, (1998), http://crdlx5.yerphi.am/proj/ani.
14. Edwards, W., Lindman, H., Savage, L.J., Bayesian Statistical Inference, in Robustness of bayesian analyses, Elsevier Science Publishers, (1984).
15. Heck, D., et.al, FZKA-Report 6019, Forschungszentrum Karlsruhe, Germany, (1998).
16. Aharonyan, F.A. , Chilingarian, A.A., et al, NIM, (1991), A-302, 522.
17. Devroye, L., Gyorfi, L., Nonparametric density estimation, The L1 view. Jown Wiley and Sons, New York, (1985).
18. Parzen, E., Ann. Math. Stat., 33, (1962), 1065.
19. Lofsgaarden, D.O., and Quesenberry, C.D., Math. Stat.,36, (1965), 1049.
20. Tapia, R.A., Thompson, J.R., Nonparametric probability density estimation, The John Hopkins University Press, Baitimore and London, (1978).
21. Chilingarian, A.A., and Galfayan, S.KH., Statet. Problems of Control, Vilnius, 66, (1984), 66.
22. Efrom, B., Canadian J. of Statist., 9, (1981), 139.
23. Toussaint, G.T., IEEE Trans. On Information, IT-20, (1974), 472.
24. Snappin, S.M., Knoke, J.D, Technometrics, 26, (1984), 371.
25. Hecht-Nielsen, R.. Neurocomputing, Addison-Westly Publishing Company, Reding, MA, (1990).
26. Petersen, C., Rognvaldsson, T., Lonnblad, L., LU TP 93-29, CERN-TH.7135/94.
27. Sobol, I.M., SIAM J. Number. Anal. 16 , (1979), 790.
28. Antoni, T., et al, J. Phys. G, 25, (1999), 2161.