PROCEEDINGS OF SCIENCE

PDE-Foam - probability-density estimation using self-adapting phase-space binning

Dominik Dannheim[*]
CERN, Switzerland
E-mail: `dominik.dannheim@cern.ch`

Probability-Density Estimation (PDE) is a multivariate discrimination technique based on sampling signal and background densities defined by event samples from data or Monte-Carlo (MC) simulations in a multi-dimensional phase space. In this paper, we discuss an innovative improvement of the PDE method that uses a self-adapting binning method to divide the multi-dimensional phase space in a finite number of hyper-rectangles (cells). The binned density information is stored in binary trees, allowing for a very fast and memory-efficient classification of events. The implementation of the binning algorithm (PDE-Foam) is based on the MC event-generation package Foam. We present performance results for representative examples (toy models). The new PDE-Foam shows improved classification capability for small training samples and reduced classification time compared to a previous PDE implementation based on range searching (PDE-RS).

XII Advanced Computing and Analysis Techniques in Physics Research
November 3-7 2008
Erice, Italy

[*] Speaker.

PoS(ACAT08)064

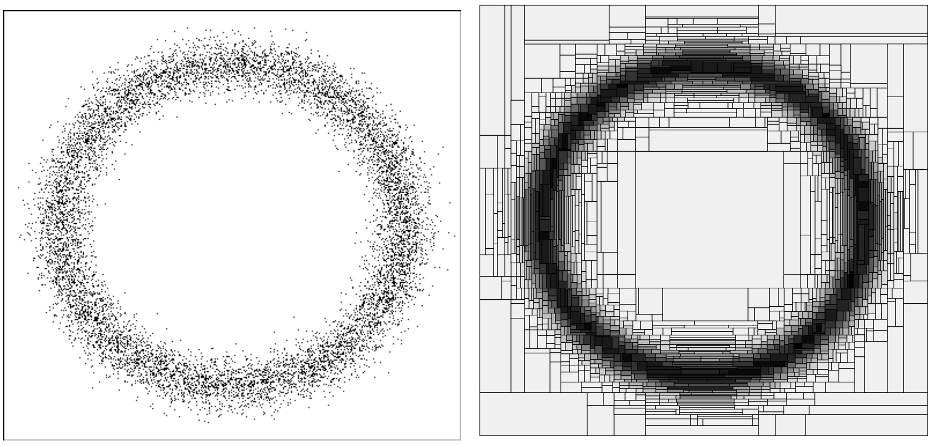
1. Probability-Density Estimation

Multivariate discrimination techniques are used in High Energy Physics to distinguish signal from background events based on a set of measured characteristic observables. Besides other sophisticated approaches, methods based on probability density estimation (PDE) are widely used. The information contained in the individual observables is combined into a single "discriminant" variable D. The value of D for a given event is based on the density of signal and background training events in the vicinity of its coordinate in the multi-dimensional phase space. Applying a cut on D allows to separate signal from background. A PDE method based on range-searching (PDE-RS) [1] has been used successfully for classification problems in higher-dimensional observable spaces and with arbitrary correlations between the observables. Large samples of MC simulated signal and background training events are stored in binary-search trees. An efficient range-searching algorithm is used to sample the signal and background densities in small multi-dimensional boxes around the phase-space points to be classified. PDE-RS shows a discrimination power similar to artifical neural networks (NNs) at largely reduced training time. It has the further advantage of transparently handling the involved statistical uncertainties and has the size of the sampling box as the only free parameter. An apparent limitation of PDE-RS, on the other hand, is the fact that large signal- and background training samples are required to densely populate the multi-dimensional phase space. Furthermore, these samples have to be accessible in the main memory of the computer used for the classification and the classification time scales with the number of training events like $T_{class} \propto N_{train} \cdot \log N_{train}$.

2. Adaptive phase-space binning with Foam

In the following we discuss an alternative method to calculate the discriminant $D(\mathbf{x})$ based on histogrammed sampling of the phase space. Only the binned density information is preserved in binary trees after the training phase, allowing for a very fast and memory-efficient classification of events.

A self-adaptive binning method, called "PDE-Foam" [2], is used to project the information contained in the signal and background samples into a grid of d-dimensional cells with non-equidistant cell boundaries, called the "foam of cells". The implementation of PDE-Foam is based on the MC event-generation package Foam [3] included in the analysis package ROOT [4] and has been developed within the framework of the Toolkit for Multivariate Data Analysis with ROOT (TMVA) [5]. The foam is iteratively produced using a binary-split algorithm for the cells acting on samplings of the input distribution within the cell boundaries. For each cell, random samplings of the input distribution are projected onto the d axes and the relative variance of the projected distributions is evaluated along the axes. The cell to be split next and the corresponding division axis and point for the split are selected as the ones for which the splitting leads to largest reduction in relative variance. After the split, the two new daughter cells become 'active' cells and the old mother cell remains in the binary tree, marked as being 'inactive'. The final number of active cells is a predefined free parameter and only limited by the amount of available computer memory. In the context of PDE, Foam has been adapted such that the splitting of cells is based on an input distribution that is sampled from MC training events using the PDE-RS method. A detailed description of the splitting algorithm can be found elsewhere [2, 3]. The geometry of the final foam reflects the distribution of the training sample: Phase-space regions where the density is approximately constant are combined in large cells, while in regions with large gradients in density many small cells are created. Figure 1(a) shows a 2-dimensional Gaussian-ring distribution[1] and Fig. 1(b) shows a graphical representation of the resulting foam with 2000 active cells. Each cell contains the number of events from the input distribution belonging to the volume of the cell. The foam consists of only a few large and sparsely populated cells in the center and corner regions of the 2-dimensional plane, where the gradient of the Gaussian radial component of the distribution is small. Close to the center of the ring, however, where the radial component of the distribution has a steep gradient, the foam consists of many small and densely populated cells.



(a) Input density (b) Foam representation

Figure 1: (a) 2-dimensional Gaussian-ring distribution sampled from 500000 events. (b) Foam representation with 2000 active cells. The level of grey indicates the event density inside the corresponding cell.

3. PDE-Foam parameters

A detailed description of the parameters steering the foam buildup is given elsewhere [2]. In the following, we give a brief overview of the main parameteres and their optimisation for reaching optimal classification performance.

3.1 Size of sampling box

The size of the box used for the phase-space sampling is a common parameter of both the PDE-Foam method and the standard PDE-RS method. In case of PDE-Foam, the box size is only relevant for the density sampling during the training phase, while for PDE-RS the box size is only used for the calculation of the discriminant during the classification phase. A larger box leads to a reduced statistical uncertainty for small training samples and to smoother sampling. A smaller

[1] The definition of this Gaussian-ring distribution corresponds to the signal distribution of the example "Highly Correlated Observables" defined and discussed in [1]. The events are distributed uniformly in the azimuth angle and according to a Gauss distribution in the radial coordinate, with mean radius of 3 and width of 0.5.

box on the other hand increases the sensitivity to statistical fluctuations in the training samples, but for sufficiently large training samples it will result in a more precise local estimate of the sampled density.

Besides affecting the estimator performance, the box size influences the training time in case of PDE-Foam and the classification time in case of PDE-RS. A larger box increases the CPU time during sampling, due to the larger number of nodes to be considered in the binary search [1].

In general, higher dimensional problems require larger box sizes, due to the reduced average number of events per box volume.

3.2 Number of cells

The target number of cells for the final foam is the main parameter impacting the accuracy of the phase-space binning. An increased number of cells leads in general to improved performance provided that sufficiently large training samples are available. However, for an increasing number of cells with small training samples, the foam becomes more vulnerable to statistical fluctuations in the training samples in particular in less populated regions of the phase space and the performance might drop when further increasing the target number of cells (overtraining). An increased number of cells also leads to increased training time and higher memory consumption to store the foam object.

Figure 2 shows the dependence of the estimator performance as function of the number of active cells for an example with five moderately correlated observables constructed from Gaussian distributions for signal and background[2]. The two curves correspond to foams build-up from small and large training samples, respectively. The small training sample consists of 5×10^4 signal and 5×10^4 background events, whereas the large training sample contains 5×10^5 signal and 5×10^5 background events. As expected, the performance of the foams built from the large training sample exceeds the one of the foams based on the small training sample. In case of the large training sample, the performance increases over a wide range of number of cells and reaches its maximum for about 20000 cells, after which it drops due to the decrease in statistical precision resulting in overtraining. For the small training sample, the maximum is already reached for foams with approximately 5000 cells and the drop in performance afterwards is steeper.

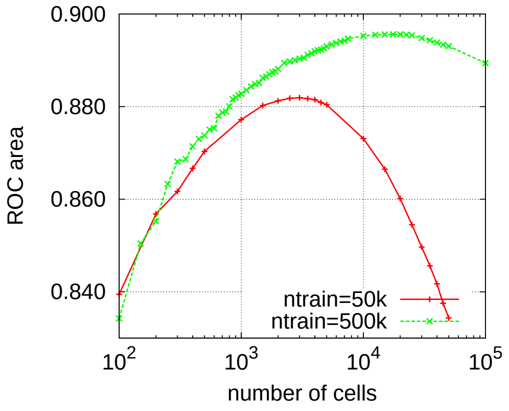


Figure 2: Dependence of the estimator performance on the number of active foam cells.

3.3 Minimum number of events N_{min}

The cell splitting algorithm assumes sufficient statistical accuracy of the sampled density distributions in all cells. This might not be guaranteed in case of small training samples, where cell

[2] The definition of the distributions corresponds to the example "High Dimensional Example" defined and discussed in [1].

splitting in scarcely populated phase-space regions can lead to overtraining effects. Therefore cells should not be taken into account for further splitting, if the number of training events contained inside a cell is too small. An adjustable parameter N_{min} has been implemented, which sets the minimum number of events contained in any cell which is considered for further splitting. If the number of events is below N_{min}, the cell is not considered for further splitting. If no more cells are available with sufficient number of events, the cell splitting stops, even if the target number of cells is not yet reached. The cut on N_{min} reduces the sensitivity to statistical fluctuations in the training samples and improves drastically the performance for small number of training events. The default value of $N_{min} = 100$ leads to a good performance for most cases studied. It can be combined with a large target number of cells, as it limits the effective number of cells sufficiently and thus avoids overtraining even for small training-sample sizes.

4. Results

In the following we present a comparison of the performance and CPU-time consumption between PDE-Foam and the standard PDE-RS method. The results are shown for the example with five moderately correlated observables. Other examples have been studied and similar results were obtained.

4.1 Performance

Figure 3 shows the estimator performance, measured by the area under the ROC curve, as function of the number of signal and background training events for foams of 1000 and 20000 active cells, respectively. Note that the number of training events corresponds to the indiviual sizes of both the signal and background samples. The actual total sample size is therefore twice the number of events shown on the x-axis. The performance of the standard PDE-RS method is also shown. Single foams were built for these examples with 2000 samplings, a sampling-box size of 0.033 and a cut on the minimum number of events per cell of $N_{min} = 100$. In case of PDE-RS, the sampling-box size was 1.2 in units of the original observables, corresponding to approximately 0.12 in normalised coordinates. For small training samples up to approximately 10^5 events, the foams perform better than the standard PDE-RS method. Apparently the geometry of the foams is well adapted to the event distributions and the implicit averaging of the event densities over the cell volumes leads to better performance than the sampling with fixed box size performed by the original PDE-RS method. For very small training samples of 30000 events and less, the foams with 1000 and 20000 cells behave identically, since the cut on the minimum number of events per cell of 100 limits the effective number of final cells to a value below 1000. For large training

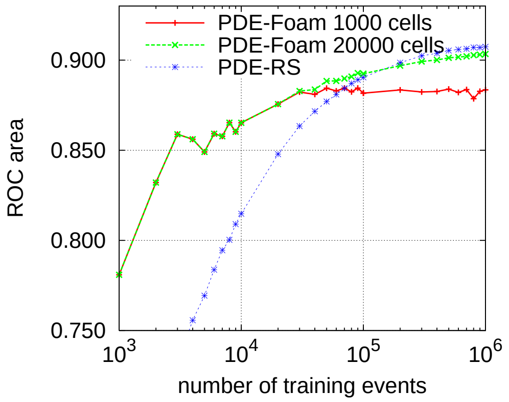


Figure 3: Estimator performance as function of the number of training events for foams with 1000 and 20000 active cells and for the standard PDE-RS method.

samples above 40000 events, the foam with 20000 cells performs better than the one with 1000, taking advantage of its finer granularity and the increased statistical precision of the larger training samples. However, for training-sample sizes of more than 2×10^5 events, it does not quite reach the performance of the standard PDE-RS method. For such large sample sizes, the local density estimates obtained with the PDE-RS method by counting events in the vicinity of the events to be classified are more precise than the density estimates from counting events in foam cells of finite granularity.

4.2 CPU time

For PDE-RS, the training time consists only of the creation of the binary search trees used to store the training samples. For PDE-Foam on the other hand, the training time is dominated by the repeated density sampling during the iterative build-up of the foam structure. On a 2.33 GHz CPU, the total training time for PDE-RS was found to be about 7 seconds for a model with 5 observables and training samples of 10^6 signal and background events each. The corresponding training time for PDE-Foam with 1000 (20000) active cells was about 4.5 (120) minutes.

For PDE-Foam, the classification time depends mostly on the number of cells in the final foam and is almost independent of the number of training events. For the standard PDE-RS method, on the other hand, the classificiation time rises with the number of training events, due to the larger size of the binary trees. With the training parameters described above, the classification time for PDE-RS, using signal and background testing samples of 5×10^5 events each, was found to be about 40 minutes. The corresponding testing time for PDE-Foam with 1000 (20000) active cells was only about 1.7 (2.7) minutes.

5. Conclusions

A new method for multivariate analysis, PDE-Foam, has been developed. It combines the adaptive binning algorithm of the Foam method so far only used for Monte-Carlo event generation with probability-density estimation based on range searching (PDE-RS). PDE-Foam has been implemented within the TMVA package for multivariate analysis. The performance of PDE-Foam exceeds the classification performance of PDE-RS for small training samples. Furthermore, it leads to largely reduced classification time. The classification time is independent of the number of training events. The main limitations of PDE-RS have therefore been overcome.

References

[1] T. Carli and B. Koblitz, *A Multi-variate Discrimination Technique Based on Range-Searching,* Nucl. Inst. Meth. A501 (2003) 576.

[2] T. Carli et al., *PDE-Foam - a probability-density estimation method using self-adapting phase-space binning*, CERN-PH-EP/2008-021, available as arXiv:physics.data-an:0812.0922.

[3] S. Jadach, *Foam: Multi-Dimensional General Purpose Monte Carlo Generator With Self-Adapting Simplical Grid*, Comput. Phys. Commun. 130 (2000) 244, physics/9910004; S. Jadach, *Foam: A General-Purpose Cellular Monte Carlo Event Generator*, CERN-TH/2002-059, physics/0203033.

[4] R. Brun and F. Rademakers, *ROOT - An Object Oriented Data Analysis Framework*, Proceedings AIHENP'96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. Meth. in Phys. Res. A 389 (1997) 81-86. See also http://root.cern.ch/.

[5] A. Höcker et al., *TMVA - Toolkit for Multivariate Data Analysis*, CERN-OPEN-2007-007 (2007), arXiv:physics/0703039v4, see also http://tmva.sourceforge.net/.