technische universität dortmund

SFB 876 Providing Information by Resource-Constrained Data Analysis

Mining for Spectra – The Dortmund Spectrum Estimation Algorithm

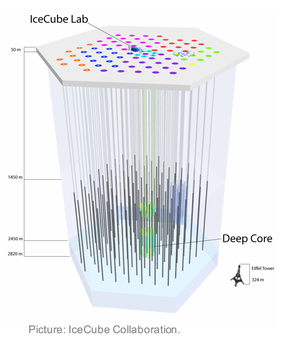
Tim Ruhe[1], Tobias Voigt[2], Max Wornowizki[2], Mathis Börner[1], Wolfgang Rhode[1], Katharina Morik[3]

[1] Lehrstuhl Experimentelle Physik 5, Department of Physics, Technische Universität Dortmund, 44221 Dortmund
[2] Lehrstuhl Statistik in den Biowissenschaften, Department of Statistics, Technische Universität Dortmund, 44221 Dortmund
[3] Lehrstuhl für Künstliche Intelligenz, Department of Computer Science, Technische Universität Dortmund, 44221 Dortmund

Algorithm

Inverse Problems in Gamma- and Neutrino Astronomy

Obtaining spectra of incident particles, such as gamma-rays or neutrinos is a common challenge in Air-Cherenkov and neutrino-astronomy.



Picture: IceCube Collaboration.


Picture: Jens Buß.

The energy of the primaries cannot be accessed directly, but has to be inferred from other observables, e.g. energy losses of secondary particles. Mathematically, this corresponds to a Fredholm integral equation of the first kind:

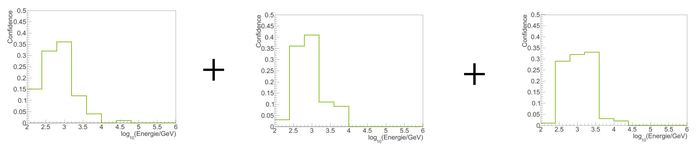
$$g(y) = \int_a^b A(E,y) f(E) dE$$

Several algorithms for the solution of this problem exist, which are, however, somewhat limited e.g. in the number of input variables, or in the sense, that information on individual events is lost.

The **Dortmund Spectrum Estimation Algorithm (DSEA)** aims at overcoming these limitations by using state of the art data mining techniques.

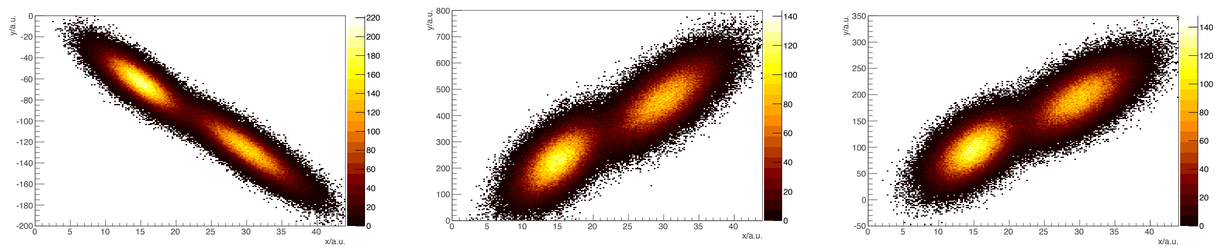
The Dortmund Spectrum Estimation Algorithm - DSEA

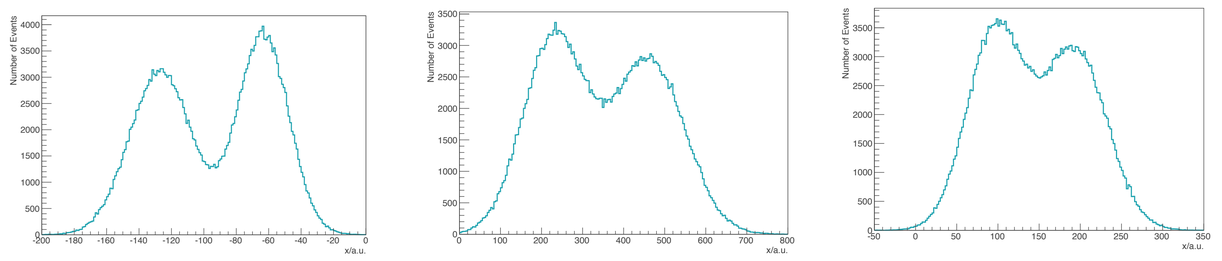
1. **Discretize** $f(x) \mapsto \vec{f}(x) = (f_1, \ldots, f_m)$. **(Initialize)**
2. **Train Model** A subset of n examples $(\underline{A}, W, L) = \{(\vec{a}, w, l)_1; \ldots; (\vec{a}, w, l)_n\}$ is used to train a model $M(\underline{A}, W, L)$. Each example consists of a label l, a weight w and h attributes $\vec{a} = (a_1, \ldots, a_h)$.
3. **Apply Model** The Model $M(\underline{A}, W, L)$ is applied to a set off $\tilde{n}$ unlabeled examples $\underline{\tilde{A}} = (\vec{\tilde{a}}_1, \ldots, \vec{\tilde{a}}_{\tilde{n}})$ yielding a confidence $c_{i,j} = g(M(\underline{A}, W, L), \vec{\tilde{a}}_i)$ for the i-th example to belong to the j-th bin in $\vec{f}(x)$.
4. **Reconstruct Spectrum** For the k-th iteration, the bin content $\hat{f}_{j,k}$ of the j-th bin is estimated as $\hat{f}_{j,k} = \sum_{i=1}^{\tilde{n}} c_{i,j}$.
5. **Update weights** The example weights for the $(k+1)$-th iteration are updated according to $w_{i,k+1} = \frac{\hat{f}_{k,l_i}}{\tilde{n}}$. **(Continue with Step 2)**

$f_j =$  $+ \cdots$

Convergence

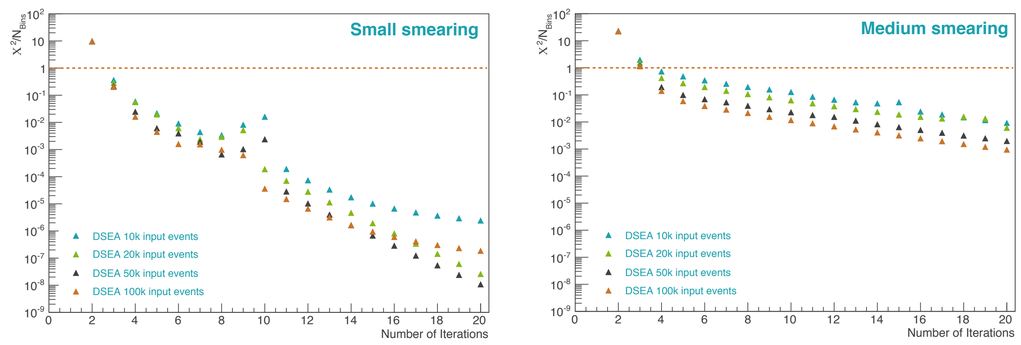
Toy Monte Carlo Simulation

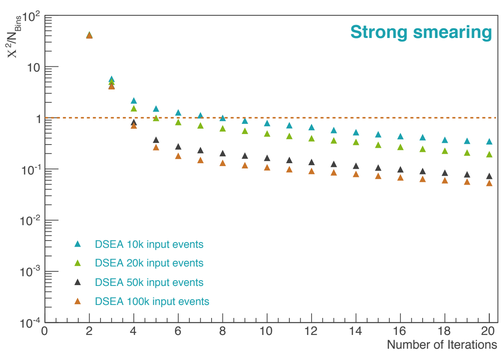




Distribution of selected input features (bottom) as well as their correlation to the class variable (top). From left to right increasing levels of smearing (small, medium, strong) are shown. In total ten attributes were generated using Gaussian smearing of the sought after variable x and utilized in the unfolding process. The level of smearing was randomly chosen from a fixed and predefined range and kept constant for every set of simulated events.

Convergence



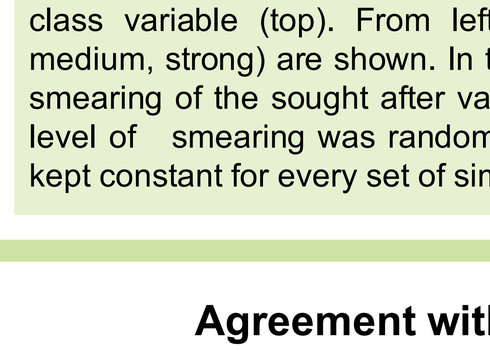


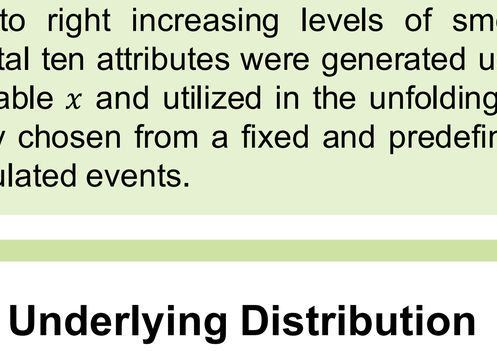
Convergence of DSEA for various levels of smearing and different numbers of input events. One finds that DSEA converges faster for smaller smearing and if given more input events. In all cases the convergence criterion (dashed line) is met after three to five iterations.

Performance

Agreement with Underlying Distribution

Agreement with the underlying pdf for different levels of smearing and different numbers of input events, evaluated using the Hellinger distance. Compared to a Random Forest regression, better agreement is obtained with DSEA, if the number of input events is sufficiently large.





Reconstructed Spectra

Reconstructed spectra obtained with DSEA for three different levels of smearing. All reconstructions were carried out using 10 attributes and 100k examples. For all three levels of smearing the reconstructed spectra were found to agree with the underlying distribution of events within the uncertainties obtained from a tenfold bootstrap.

Lehrstuhl für Experimentelle Physik 5

Lehrstuhl Statistik in den Biowissenschaften

Lehrstuhl für Künstliche Intelligenz, LS 8