PROCEEDINGS OF SCIENCE

Summary of Session 2: Data Analysis - Algorithms and Tools

Thomas Speer[*]
Physik-Institut der Universität Zürich,
Winterthurerstrasse 190, CH-8057 Zürich, Switzerland.
E-mail: `Thomas.Speer@cern.ch`

This report gives a short overview of the main topics and presentations given in Session 2. This session is devoted to new ideas and development of algorithms, statistical methods and tools for data analysis, reconstruction and visualization in nuclear and particle physics.

XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research
April 23-27 2007
Amsterdam, the Netherlands

[*] Speaker.

© Copyright owned by the author(s) under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike Licence. http://pos.sissa.it/

1. Introduction

The main emphasis of this session is on new ideas and development of algorithms, statistical methods and tools for data analysis, reconstruction and visualization in nuclear and particle physics. The large number of contributions (42) presented and their variety is indicative of the breadth of the session and the activity in these fields. Several contributions were also held in common with Session 1 or Session 3, and a common round-table discussion was held with Session 1.

2. Presentations

2.1 Multi-variate analysis methods

A large fraction of the contributions were related to multi-variate analysis methods, describing both novel algorithms and uses of established techniques. Their applications include trigger and event selection, reconstruction and data analysis, in a variety of experiments, both in high-energy physics and astro-particle physics. A historical perspective on the usage of multi-variate analysis methods in high-energy physics was given by P. Bhat, with a review of the first results using adaptive methods and the evolution of their usage in the two experiments at the Tevatron collider at Fermilab, CDF and D0.

A welcome effort is the development of generic multi-purpose multi-variate analysis packages specifically geared towards high-energy physics applications. Two such packages have been presented, *TMVA* (H. Voss) and *StatPatternRecognition* (A. Buckley). These projects provide standardized implementations of several multi-variate analysis methods, with a common platform and interface for the implemented classifiers. In addition, they provide tools for their training, testing and evaluation, with visualization tools and easy-to-use graphic user interfaces. This will ease to the use of multi-variate analysis methods and allow for a consistent and objective evaluation of the different algorithms, since these can be trained and tested on the same data-sample. Both packages are free and open-source, which will allow inspection of the code, and addition of new methods and tools by users. Furthermore, the redundancy provided by the implementation of several popular algorithms in both packages will help to compare and verify the implementations.

Several contributions presented applications to high-energy physics of of previously seldomly-used methods. The *support vector machine* algorithm has been presented in the context of τ-tagging in the ATLAS experiment at the LHC, and compared to other methods (M. Wolter). The basic idea of this method is to build a hyperplane that separates signal and background vectors (events) using only a minimal set from the training sample (support vectors). The position of the hyperplane is obtained by maximizing the margin between it and the support vectors. The performance of this method has then been compared to other multi-variate analysis methods and a traditional cut-based method. While first result show, as expected, a clear improvement of the multi-variate analysis methods used over the cut-based method, little difference can be seen between the different adaptive methods themselves.

Similarly, the *self-organizing maps* algorithm has been presented in the context of b-tagging in the CMS experiment at the LHC (A. Heikkinen). Self-organizing maps allow the mapping from an n-dimensional input data space onto a regular two-dimensional array of neurons, such that every neuron of the map is associated with an n-dimensional reference vector. The neurons of the map

PoS(ACAT)089

are connected to adjacent neurons by a neighborhood relation, which dictates the topology of the map, and similar input patterns are mapped to adjacent regions of the characteristics map. During the unsupervised training phase, the map forms an elastic net that folds onto the cloud formed by the input data and approximates the density of the data.

The use of this method to identify electrons and jets based on calorimetric information in the high-level trigger of the ATLAS experiment was also shown (J. Seixas). Here, a *learning vector quantization* algorithm improves the classification performance of self-organizing maps, and the hypothesis testing procedure was performed by a *multi-layer perceptron* neural network. In an alternative method (R.C. Torres), a neural-network classifier is used after a pre-processing stage, where a non-linear decorrelation of the energy sum of rings around the most energetic calorimeter cell is preformed in an independent component analysis. The results show that both methods achieve a better performance with respect to the baseline algorithm, which uses basic clustering strategies and applies linear cuts, and that they can be applied within the time-constraints of the high-level trigger.

Evolutionary algorithms were featured in two contributions. Natural evolution can broadly be summarized as the mechanism to generate a population of individuals with increasing fitness to their environment. Evolutionary computation attempts to simulate natural evolution on a computer to generate a set of solutions of increasing quality. This can be achieved through processes leading to maintenance or increase of a population, their ability to survive and reproduce in a specific environment, and an estimator that quantitatively measures the evolutionary fitness. Each candidate solution is encoded in a chromosome made of constituent genes, and a set of genetic operators simulates the reproduction process, introducing thus genetic variation. By selecting for reproduction the candidates with the highest fitness estimate, the quality of the solutions is improved though successive iterations. The main evolutionary algorithms, *genetic algorithms*, *genetic programming* and *gene expression programming* differ mainly by the gene encoding and reproduction methods employed.

The *gene expression programming* algorithm and a first application to high-energy physics were shown by L. Teodorescu. The main feature of this algorithm is that, in addition to the chromosome, it encodes each candidate solution in a mathematical expression (called *expression tree*). A first application of this algorithm applied to optimize the selection of $K_s \rightarrow \pi^+\pi^-$ decays was shown. A package that implements a *genetic algorithm* for rectangular cut optimization for physics analysis (GARCON) was shown by A. Drozdetskiy. This package allows an automatic optimization and verification for stability in a multidimensional phase space. It is a publicly available tool, which has already been used in several Monte Carlo physics studies in the CMS experiment.

Two contributions presented applications of neural networks in astro-particle physics experiment. In the first (S. Riggi), neural networks are studied to determine the mass composition of ultra-high energy cosmic rays at the Pierre Auger experiment. This would allow to study possible correlations between the mass of the incoming cosmic ray and the arrival direction of the shower at the ground and to correct the reconstructed energy of the shower with a missing energy factor. In the second contribution (S. Khatchadourian), neural networks are studied for the trigger of the HESS 2 experiment, with the aim of distinguishing the extensive air showers created by high-energy gamma rays from those created by particles such as muons or protons. An alternative method to detect cosmic rays using the front scattering of electromagnetic waves on ionized atoms

PoS(ACAT)089

in the atmosphere by the shower was also presented (L. Andrade). A matched filter is then used to maximize the signal-to-noise of the detected signal.

2.2 Reconstruction algorithms and advanced data analysis environments

Several recent developments were presented. V. Kartvelishvili presented a novel electron bremsstrahlung recovery algorithm in the context of the ATLAS silicon tracker, using dynamic noise adjustment. This algorithm attempts to find layers in the tracker where bremsstrahlung occurred along the trajectory of the track, and the fraction of energy retained by the electron is estimated. An effective noise matching the probability of the estimated retained energy is then calculated by mapping the cumulative Bethe-Heitler distribution onto a unit Gaussian distribution. This effective noise term is then used in a simple Kalman filter, in the same way as multiple scattering or other noise terms are incorporated during filtering. Studies have shown that this algorithm yields improved residual distributions, both in terms of width and bias, and improved error estimates with respect to the default Kalman filter. In another presentation, a fast vertex fitter, developed in the context of the ATLAS high-level trigger, was presented (D. Emeliyanov). In this algorithm, the track covariance matrices are used directly, without requiring computing-intensive matrix inversions.

Other contributions dealt with energy loss corrections in the calorimeter of the BES III experiment (M. He), the matching between reconstructed tracks and clusters in the electromagnetic calorimeter of the ALICE detector (A. Pulvirenti), the alignment of the ATLAS inner detector (R. Haertel) and a method to determine the luminosity spectrum of a future high luminosity electron-position linear collider using Bhabha events was presented (A. Shibata).

Trigger algorithms were also featured in several presentations. A trigger to select cosmic rays detected by the hadronic calorimeter of the ATLAS experiment was presented by B. Ferreira. This trigger algorithm, which has been used to commission the calorimeter, uses a whitening filter and a matched filter to improve its performance. In addition, the muon trigger of the ATLAS experiment was presented by C. Siragusa, and the Global Tracking Trigger of the ZEUS experiment by V. Roberfroid.

Finally, several presentations described data analysis environments and frameworks. A reconstruction tools for the study of short-lived resonances in the ALICE experiment was presented by F. Riggi. The reconstruction of short-lived resonances requires optimal performance of the complete reconstruction chain, from the reconstruction of primary vertices to track reconstruction and particle identification. In addition, one contribution described a framework for the reconstruction and analysis of emulsion data as used at the OPERA neutrino experiment (V. Tioukov), and one contribution described a framework dedicated to *B*-physics Analysis at the ATLAS experiment (J. Catmore).

2.3 Statistical methods

Several novel statistical methods, for a wide range of applications, were presented in this workshop. These include a method to calculate an upper limit for Poisson variables incorporating systematic uncertainties using a Bayesian approach (Y. Zhu), a modified Pearson-χ^2 test to compare histograms (N. Gagunashvili), a two-dimensional Kolmogorov-Smirnov test (R. Lopes), and

PoS(ACAT)089

a procedure for unfolding distributions from experimental data using machine learning methods (N. Gagunashvili).

3. Conclusion

With a large fraction of the contributions related to multi-variate analysis methods, the round-table discussion addressed issues facing multi-variate analysis methods and results based on these methods. During the discussion, the need to generate a methodology to study multi-variate analysis methods and have common benchmark problems on which these methods could be compared has been highlighted. The need of standardized, multi-purpose implementations was also stressed, which is starting to be addressed by the *TMVA* and *StatPatternRecognition* presented in this workshop.

Once again, this workshop has proven to be an excellent forum for discussions and valuable exchanges of ideas on a wide range of topics related to data analysis methods and algorithms.