International Journal of Modern Physics A,

NEW DIRECTIONS IN DATA ANALYSIS

PUSHPALATHA C. BHAT
Fermi National Accelerator Laboratory, P.O. Box 500*
Batavia, Illinois 60510, USA

In the next decade, high energy physicists will use very sophisticated equipment to record unprecedented amounts of data in the hope of making major advances in our understanding of particle phenomena. Some of the signals of new physics will be small, and the use of advanced analysis techniques will be crucial for optimizing signal to noise ratio. I will discuss new directions in data analysis and some novel methods that could prove to be particularly valuable for finding evidence of any new physics, for improving precision measurements and for exploring parameter spaces of theoretical models.

1. Introduction

We are building powerful accelerators and sophisticated detectors in the hope of making major discoveries in the next decade. We hope to discover the Higgs boson, Supersymmetry or Technicolor or something completely unexpected and exalting! We would also like to make precision measurements of some of nature's fundamental parameters. In order to achieve these goals it is crucial that we employ advanced and optimal data analysis methods both on-line and off-line.

In the not so distant past, we could afford the luxury of writing data onto storage media with simple interaction triggers and organize, reduce and analyze data completely off-line. But, as we learnt more about the world and began to address more complex problems, looking for extremely rare processes at higher beam energies and higher luminosities, we had to handle and sift through large amounts of data on-line before selected data are written out. The new generation of experiments will be a lot more demanding than the previous in data handling at all stages; the rates of interactions to be handled and the number of detector channels read-out will grow by orders of magnitude. Finding the signals of new physics becomes a veritable case of "finding needles in a hay-stack". The unprecedented challenges will require new paradigms and technologies to be identified, developed and adopted.

2. Intelligent Detectors and Smart Triggers

The data analysis in HEP experiments starts when a high energy event occurs. The data from the detectors must be transformed into useful "physics" information in real-time. The calorimeter, for instance, can have "intelligence" close to its electron-

*Operated by the Universities Research Association, Inc., under contract with the U. S. Department of Energy.

ic read-out so that the clustering and energy measurements are readily available. Such information from different sub-detectors can be used to extract event features, such as the number of tracks, high transverse momentum (p_T) objects and object identities. These features can then be used to make a global decision about whether or not the event is potentially interesting. Therefore, we need to build *intelligent detectors* and *triggers*. The feature extraction and further processing such as particle identification or event classification can be accomplished using smart algorithms either built into hardware (neural networks chips, for example) or configured in generic hardware such as Field Programmable Gate Arrays (FPGAs) or Digital Signal Processors (DSPs). The H1 experiment at HERA has implemented neural network hardware in its Level-2 trigger[1] which has operated successfully since 1996 and has been crucial for the rich harvest of physics results from H1. Innovative data management on-line, using for example RAM disks, and employing algorithms in trigger hardware would be beneficial in meeting the demands of data handling and analysis on-line. Use of expert or fuzzy-logic systems in controls and monitoring of detector electronics is an area that needs to be explored as well.

3. Optimal Analysis Methods

When classifying events, the traditional procedure of choosing and applying cuts on one event variable at a time is rarely optimal in the sense of minimizing the probability to mis-classify events. By contrast, given a set of event variables, that, in general, are correlated, optimal separation can always be achieved if one treats the variables in a fully *multivariate* manner. Some classes of data analysis tasks that benefit from multivariate methods are particle identification (electrons, taus, b-quark jets etc., γ/π^0 separation, quark/gluon jet separation), signal-background discrimination, parameter estimation (tracking, vertexing, mass measurements etc), function approximation (energy correction functions, mis-identification rates) and data exploration (latent structure analysis, multivariate bump hunting).

There are a number of parametric and non-parametric multivariate methods from the simplest Fisher linear discriminant to the sophisticated non-linear, adaptive methods[2]. Neural networks have emerged as powerful and flexible methods of multivariate data analysis. These and other multivariate methods have been used in recent experimental data analyses around the world: DØ, CDF at the Tevatron, LEP experiments at CERN, experiments at DESY and SLC. These will be the methods of choice for future analyses.

4. Some Examples

Because of a lack of space, I describe a couple of example analyses from the DØ experiment and a Monte Carlo study (with apologies to other experiments).

The top quark mass measurement was one of the most important results from the last run of the Tevatron experiments[3]. Since the DØ experiment did not have a silicon vertex detector (SVX) in Run I and used only soft muon tagging for b-jet identification, the b-tagging efficiency was only 20% in the lepton + $\geq$ 4jets channel compared to approximately 53% at CDF which had the ability to tag b-jets with its SVX. Nonetheless, DØ was able to measure the top quark mass with a precision approaching that of CDF, by using multivariate techniques for separating signal and background while minimizing the correlation of the selection with the top quark mass. Two multivariate methods, (1) a log-likelihood technique and (2) a feed forward neural network, were used to compute a signal probability $P(top \mid D)$ for each event, given data D. A likelihood fit (based on a Bayesian method[4]) of the data to a discrete set of signal and background models in the $[P(top \mid D), m_{fit}]$ plane was used to extract the top quark mass. Combining results from the two methods, taking into account their correlation, an overall measurement of $m_t = 173.3 \pm 7.8$ GeV/c^2 was obtained. The top quark mass measurement in the dilepton+jet channel also benefitted from a multivariate method (probability density estimator).

Multivariate analysis methods enabled DØ to establish the world's best limits[5] on the existence of leptoquarks that might decay into electrons and quarks. The lower limit on the leptoquark mass from the DØ experiment was 225 GeV, ruling out the possibility that the HERA event excess (reported in February 1997) can be interpreted as an evidence for first generation scalar leptoquark production.

In the next few years, a low mass Higgs boson may be discovered at the Fermilab Tevatron, a possibility that has motivated an intense study[6]. We have studied[6] the potential of the CDF and DØ experiments to make such a discovery in Run II, via the processes $p\bar{p} \to WH \to l\nu b\bar{b}$, $p\bar{p} \to ZH \to l^+l^-b\bar{b}$ and $p\bar{p} \to ZH \to \nu\bar{\nu}b\bar{b}$. We have shown that a neural network analysis could yield a 5σ discovery for $100 \leq M_H \leq 130$GeV/c^2 with only half the integrated luminosity needed for a conventional analysis. Fig. 1 shows the neural network distributions for signal monte carlo events with $M_H = 110$ GeV/c^2 compared with the specified backgrounds, for a set of seven input variables. (For details, see ref. 6). A plot of the required integrated luminosity for a 5σ observation is also shown in Fig 1.

5. Exploring Models

Physicists are becoming increasingly convinced of the value of Bayesian reasoning as a powerful way of extracting information from data and of updating our knowledge upon arrival of new data. The Bayesian approach provides a well-founded mathematical procedure to compute the conditional probability of a model and therefore to do straight-forward and meaningful model comparisons. It also allows treatment of all uncertainties in a consistent manner. Practical applications of these ideas to (1)fitting binned data to one or more multi-source models[4] and (2)the extraction of the solar neutrino survival probability[7] using data and solar neutrino model predictions illustrate the usefulness of Bayesian methods in data analysis.

This approach provides a systematic way of extracting probabilistic information for each parameter of a model, say for example a particular SUSY model, via marginalization over the remaining parameters. I believe that this probabilistic approach to model exploration could prove to be extremely fruitful. Studies of this approach are in progress.

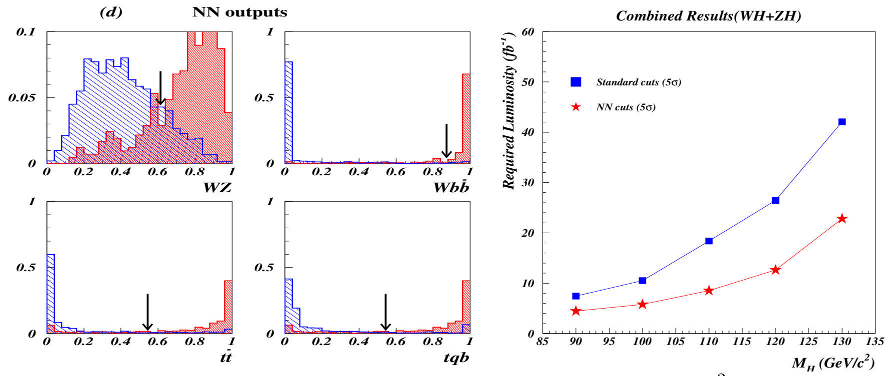

Fig. 1. (Left)The neural network distributions for WH (M_H = 110 GeV/c^2) signal Monte Carlo events (heavily shaded) compared with $Wb\bar{b}$, WZ, $t\bar{t}$, single top background events. (Right) Comparison of the required integrated luminosities for a 5σ observation in the CDF and DØ experiments with WH and ZH channels combined for NN and conventional cuts.(See ref. 6)

6. Summary

Major discoveries may await us using the next generation of experiments. Use of advanced *optimal* analysis methods will be crucial to discover and study new physics. Multivariate methods, particularly neural network techniques have already had an impact on discoveries and precision measurements and will be the methods of choice in future analyses. We have only scratched the surface in our use of powerful, multivariate methods for data exploration, visualization and statistical analysis. I believe that hybrid methods combining "intelligent" algorithms and the Bayesian/probabilistic approach will be the wave of the future.

Acknowledgements

The author would like to thank Harrison Prosper for a delightful collaboration on many of the pursuits discussed in this paper.

References

1. C. Kiesling, Proceedings of the ACAT2000 workshop, Fermilab, Batavia, IL, October 16-20, 2000, See http://conferences.fnal.gov/acat2000/.
2. P.C. Bhat *ibid.*
3. P.C. Bhat, H.B. Prosper and S.S. Snyder, *Int. J. Mod. Phys.* Vol. 13 (1998) 5113;
   DØ Collaboration, *Phys. Rev.* D Vol. 58 (1998) 052001;
   CDF Collaboration, *hep-ex/0006028*.
4. P.C. Bhat, H.B. Prosper and S.S. Snyder, *Phys. Lett.* B Vol. 407, (1997) 73.
5. DØ Collaboration, *Phys. Rev. Lett.* Vol. 80 (1998) 2051.
6. Run II Higgs Working Group Report, *hep-ph/0010338*;
   P.C. Bhat, R. Gilmartin and H.B. Prosper, *Phys. Rev.* D Vol. 62 (2000) 074022.
7. C.M. Bhat *et. al.*, *Phys. Rev. Lett.* Vol. 81 (1998) 5056.