SIGNAL ENHANCEMENT USING MULTIVARIATE CLASSIFICATION TECHNIQUES AND PHYSICAL CONSTRAINTS

R. VILALTA and P. SARDA

Dept. of Computer Science, University of Houston, 4800 Calhoun Rd., Houston TX 77204, USA
E-mail: {vilalta,ppsarda}@cs.uh.edu

G. MUTCHLER and B. P. PADLEY

Bonner Nuclear Lab, Rice University, 6100 Main Street, Houston, TX 77005, USA
E-mail: {mutchler,padley}@rice.edu

S. TAYLOR

Dept. of Physics and Astronomy, Ohio University, 251 Clippinger Labs, Athens, OH 45701, USA
E-mail: staylor@jlab.org

We report on an empirical comparison of several multivariate classification techniques (e.g., random forests, Bayesian classification, support vector machines) for signal identification; our experiments use K* mass as a test case. We show 1) the effect of using different cost matrices in generalization performance and 2) how information about physical constraints obtained from kinematic fitting procedures can be used to enrich the original feature representation. The latter step is done through a derivation of Λ particle parameters (e.g., momentum, energy, and mass) using kinematic fitting; the degree of fit using a χ^2 statistic is used as a new feature. Overall, our goal is to investigate how to incorporate physical constraints to improve classification performance.

1. Introduction

The purpose of this analysis is to gain insight on how to exploit multivariate techniques and physical constraints for signal classification and enhancement. Traditional techniques that exploit physical constraints use "kinematic fitting" to improve measured quantities and to provide a means to cut background. We propose an additional step where a multivariate classification technique is invoked on Monte Carlo data to generate a predictive model. The model is used to separate signal events from background events. Applying the model to real data results in a (predicted) signal distribution where evidence for the existence of a particle of interest is enhanced.

1.1. *The Physical Experiment*

We begin by describing the physical experiment. A broad band energetic photon beam (γ) hits a liquid hydrogen target, the proton (p). The photon interacts and produces a number of charged and uncharged particles. We will look for the following reaction:

$$\gamma \mathrm{p} \rightarrow \Lambda K^{*+} \tag{1}$$

$$\gamma \mathrm{p} \rightarrow \Lambda K^{+} \pi^{0} \tag{2}$$

$$\gamma \mathrm{p} \rightarrow K^{+} \mathrm{p} \pi^{-} \pi^{0} \tag{3}$$

Our data set contains information about the incident photon (γ), and three charged particles, K^+, p, and π^-. While the charged particles are detected, the uncharged ones are not seen, and must be inferred from the missing mass (e.g., π^0).

For each detected charged particle we measure the momentum p and the polar angle θ and azimuthal angle ϕ. From these quantities we can construct the three vector, $\mathbf{p} = \mathbf{i}p_x + \mathbf{j}p_y + \mathbf{k}p_z$ where $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are the unit vectors. We also measure the Time-of-Flight (TOF). From the TOF and momentum we can calculate the mass m of the particle. Finally, for each particle, we are able to construct a 4-vector, $(E, \mathbf{p})$, where $E = \sqrt{p^2 + m^2}$.

In this particular paper we focus on identifying the presence of K^{*+} after the photon-proton interaction (γp). This is in practice not of real interest, but stands as a convenient test case to assess the value behind multivariate classification techniques. Invoking these techniques is justified by the inherent difficulty in separating signal events from background events (many background reactions produce similar measured particles).

1.2. *Using Kinematic Fitting and Physical Constraints*

At first we applied the technique of kinematic fitting [1]. This technique takes advantage of constraints such as energy and momentum conservation to improve measured quantities and to provide a means to cut background. We have chosen to use the Lagrange multiplier method. First, the unknown variables are divided into a set of measured variables ($\vec{\eta}$) and a set of unmeasured variables ($\vec{\xi}$) such as the missing momentum or the 4-vector for a decay particle. For each constraint equation a new variable λ_i is introduced. These variables are the Lagrange multipliers. To find the best fit we minimize

$$\chi^2(\vec{\eta},\vec{\xi},\vec{\lambda}) = (\vec{\eta}_0 - \vec{\eta})^T V^{-1}(\vec{\eta}_0 - \vec{\eta}) + 2\vec{\lambda}^T\vec{f} \quad (4)$$

by differentiating χ^2 with respect to all the variables, linearizing the constraint equations and iterating. Here $\vec{\eta}_0$ is a vector containing the initial guesses for the measured quantities, V is the covariance matrix comprising the estimated errors on the measured quantities, and $\vec{f}$ represents the constraints such as energy and momentum conservation.

1.3. *Generating Confidence Levels*

For our purposes, we are interested in using kinematic fitting to obtain a confidence level (goodness of fit to the data). As an example, let's look into the fitting procedure as applied to the proton (p) and pi-minus (π^-) tracks with the Λ hypothesis. Explicitly, the constraint equations are as follows:

$$\vec{f} = \begin{bmatrix} E_p + E_\pi - E_\Lambda \\ \vec{p}_p + \vec{p}_\pi - \vec{p}_\Lambda \\ (y - y_\pi)p_\pi^z - (z - z_\pi)p_\pi^y \\ (x - x_\pi)p_\pi^z - (z - z_\pi)p_\pi^x \\ (y - y_p)p_p^z - (z - z_p)p_p^y \\ (x - x_p)p_p^z - (z - z_p)p_p^x \end{bmatrix} = \vec{0}. \quad (5)$$

The χ^2 distribution for this fit is the result of a fit to the histogram using the functional form of a χ^2 distribution with two degrees of freedom plus a flat background term. Explicitly,

$$f(\chi^2) = \frac{P_1}{2} e^{-P_2\chi^2/2} + P_3. \quad (6)$$

P_2 is a measure of how close the distribution in the histogram is to an ideal χ^2 distribution, for which $P_2 = 1$. The Confidence Level (CL) is the primary measure of the goodness of fit to the data and is given by the equation

$$CL = \int_{\chi^2}^{\infty} f(z;n)dz \quad (7)$$

where f(z:n) is the χ^2 probability density function with n degrees of freedom (where we have assumed normally distributed errors).

2. Using Multivariate Classification Techniques

In addition to the traditional approach of kinematic fitting, we suggest using multivariate classification techniques for signal identification and enhancement. Our approach consists of using the confidence levels (goodness of fit to the data described above) as new features into a classification problem. The resulting model implicitly uses the kinematic fitting results to further enhance the signal of interest (e.g., to enhance K^{*+}).

2.1. *The Classification Problem*

We begin by giving a brief overview of the classification problem [2, 3]. A classifier receives as input a set of training examples $T = \{(\mathbf{x}, y)\}$, where $\mathbf{x} = (a_1, a_2, \cdots, a_n)$ is a vector or point in the input space ($x \in \mathcal{X}$), and y is a point in the output space ($y \in \mathcal{Y}$). We assume T consists of independently and identically distributed (i.i.d.) examples obtained according to a fixed but unknown joint probability distribution. The outcome of the classifier is a function h (or hypothesis) mapping the input space to the output space, $h : \mathcal{X} \to \mathcal{Y}$. Function h can then be used to predict the class of previously unseen attribute vectors.

2.2. *Data for Analysis*

In our study, the output variable for each event indicates if the photon-proton interaction resulted in the production of K^{*+} (positive event) or not (negative event). Each feature vector $\mathbf{x}$ is made of 45 features. The first 4 features are confidence level numbers derived from the kinematic fits (Section 1.3). The next feature corresponds to the total energy. The last 40 features characterize 8 particles (3 of them detected and 5 inferred). Each particle is represented by energy E, momentum p, polar angle θ, azimuthal angle ϕ, and mass squared m^2.

Table 1. Columns 2-3: Mean accuracy performance (Acc.) with different misclassification costs. Numbers enclosed in parentheses represent standard deviations. Columns 4-5: Mean false positive rates (FPR) with different misclassification costs.

Analysis Technique	Acc. Equal Costs	Acc. Unequal Costs	FPR Equal Costs	FPR Unequal Costs
Naive Bayes	85.59 (0.86)	86.79* (0.78)	20.1	6.8
Support Vector Machines	87.69 (0.70)	88.29 (0.51)	18.7	1.6
Multilayer Perceptron	88.57 (0.85)	90.58 (0.73)	14.3	3.0
ADTree	88.90 (1.14)	90.81* (0.96)	11.5	3.7
Decision Tree	89.23 (0.93)	91.97* (0.87)	12.7	4.7
Random Forest	90.02 (1.12)	92.34* (0.95)	11.6	4.3

Our data set is derived using the CEBAF large angle spectrometer (CLAS). We gathered 1000 Monte Carlo signal events and 6000 Monte Carlo background events. The real data comprised about 13,500 events.

2.3. *Using Monte Carlo Data and Variable Misclassification Costs*

Our first set of experiments were limited to Monte Carlo data for which the value of the output variable of each event is known. Our study compared the performance of several classification algorithms in terms of predictive accuracy. We employed several algorithms including decision trees, support-vector machines, random forests, etc.

First we reduced the original size of the input space through a feature selection process, using information gain as the evaluation metric [3]. For each algorithm we varied the amount of misclassification costs. Table 1 shows our results. The first column describes the multivariate classification techniques used for our experiments. The second column shows accuracy estimations with equal misclassifications costs; the third column shows accuracy estimations where the cost of a false positive is 3 times more expensive than the cost of a false negative. Each result is the average of 5 trials of 10-fold cross validation each [3]. An asterisk at the top right of a number implies the difference is significant at the $p = 0.01$ level (assuming a two-tailed t-student distribution). Overall there is a significant increase in performance by adding a penalty when mislabelling background events as target events. In addition, Table 1 shows how for this particular domain, varying misclassification costs can yield a significant reduction in the false positive rate (FPR %, columns 4-5).

Our results denote a preference for the strategy behind "random forests". We have observed similar results in other experiments [4]. Random forests have the ability to reduce the variance and bias components of error by voting over multiple decision trees using on each tree a random selection of features [5]. They exhibit robust behavior against problems with multiple systematic errors as is common to problems in particle physics.

2.4. *Signal Enhancement on Real Data*

Our next set of experiments used real data for which the value of the output variable of an event is unknown. In this case the problem is not to maximize accuracy performance (i.e., minimize a risk functional such as zero-one loss) but instead to provide enough evidence to believe that the signal event occurred multiple times during the photon-proton interaction. The goal is to find a technique able to enhance the signal distribution over the background distribution.

Our approach to deal with the signal enhancement problem is as follows. Applying a multivariate technique M on Monte Carlo data yields a predictive model h_M. One can then apply h_M on the real data to generate a histogram for the predicted signal distribution. If model h_M exhibits good performance, we expect the histogram generated through h_M to provide evidence for the occurrence of the desired signal.

To illustrate our approach Figure 1 (left) shows a histogram generated with all real data; the x-axis corresponds to the squared mass (m^2) of the signal particle (K^{*+}). Figure 1 (middle) shows a histogram generated by taking only those events predicted as signal on the real data by a classification model. Kinematic fitting variables were part of the feature vectors. We employed random forests as the clas-

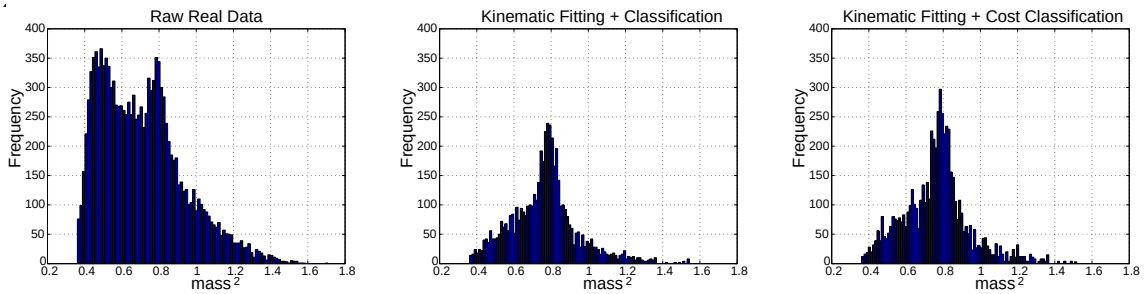

Fig. 1. Histograms using (left) real data (middle) predicted signals on real data by random forests, and (right) predicted signals on real data by random forests using cost-sensitive information. The x-axis corresponds to K^{+*} squared mass (units are in $\frac{\text{GeV}^2}{c^4}$).

sification technique; the derived information helps isolate and enhance the signal distribution. Figure 1 (right) shows the corresponding histogram using random forests with cost sensitive classification and kinematic fitting variables. The resulting histogram shows an even larger enhancement over the signal distribution.

To quantify the difference between Figure 1 (middle) and Figure 1 (right), we computed the distance between each of these empirical distributions. We used relative entropy [6] $K(f_1||f_2)$ to compute the distance between probability distributions f_1 and f_2, where

$$K(f_1||f_2) = \sum_i f_1^i \log \frac{f_1^i}{f_2^i} \tag{8}$$

and index i varies along the values of the random variable. In our case, let f_r be the distribution for the real data (Figure 1 left), f_{kc} be the distribution for kinematic fitting and classification (Figure 1 middle), and f_{kcs} be the distribution for kinematic fitting and cost sensitive classification (Figure 1 right). We found empirically the following results: $K(f_r||f_{kc}) = 0.2798$; $K(f_r||f_{kcs}) = 0.4048$. This indicates the distribution obtained by combining kinematic fitting with cost-sensitive classification yields a new signal distribution that has a larger separation from the original real data (in terms of relative-entropy).

3. Conclusions

Our study suggests generating a predictive model over Monte Carlo data to produce a distribution over real data where a signal of interest is enhanced. Our model integrates information about physical constraints using kinematic fitting.

Our current work adds confidence levels derived from kinematic fitting as new features for classification. One unexplored area is to determine the degree to which multivariate classification techniques contribute to signal enhancement without any information derived from kinematic fitting. It is important to understand how current classification techniques can exploit information derived from physical constraints.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants no. IIS-431130 and IIS-448542, and by the Department of Energy under Grant no. DE-FG03-93ER40772.

References

1. A. G. Frodesen (1979). *"Probability and Statistics in Particle Physics"*, Oxford University Press.
2. R. O. Duda, P. E. Hart, and D. G. Stork (2001). *"Pattern Classification"*, John Wiley Ed. 2nd Edition.
3. T. Hastie, R. Tibshirani, and J. Friedman (2001). *"The Elements of Statistical Learning, Data Mining, Inference, and Prediction"*, Springer-Verlag.
4. P. Bargassa, S. Herrin, S-J Lee, P. Padley, R. Vilalta (2005). *"Application of Machine Learning Tools to Particle Physics"*, Conference on Statistical Problems in Particle Physics, Astrophysics and Cosmology (PHYSTAT05).
5. L. Breiman (2001). *"Random Forests"*, Machine Learning 45(1) pp. 5-32. Springer Science-Business Media.
6. T. M. Cover and J. Thomas (1991). *"Elements of Information Theory"*, Wiley-Interscience.