

# Evaluating And Constructing Features For Identification Of Tau Leptons

R. Vilalta, A. Bagherjeiran, C. Sun

*University of Houston, 4800 Calhoun Rd., Houston TX 77204-3010, USA*

B. P. Padley, S. J. Lee

*Bonner Nuclear Lab, Rice University, 6100 Main Street, Houston, TX 77005, USA*

In this paper we show the importance of choosing the right feature representation in attempting to improve the quality of a predictive model. We explain how to evaluate and construct new features using information-theoretic measures (information gain, gain ratio) and statistical tests (e.g.,  $\chi^2$ ,  $G$  statistic). Our experiments use Monte-Carlo simulated data containing both  $\tau$  lepton signals and background events. Results show how our evaluation process can identify a small set of relevant features that bear correlation with the class ( $\tau$  signals). We also show how to construct new features by exploring the space of logical feature combinations using genetic algorithms; the set of newly constructed features can effectively improve the quality of the feature representation.

## 1. Introduction

The aim of this study is to construct good predictive models for the identification of  $\tau$  leptons; we wish to identify clusters of energy obtained from particle detectors associated with jets that are characteristic of  $\tau$  leptons (e.g., tightly collimated jets of energy). The problem is complex because the distribution of energy corresponding to  $\tau$  decay overlaps with that corresponding to the fragmentation of quarks [8]. Our approach is to use multivariate data analysis techniques for classification to separate  $\tau$  leptons from background events.

From a pattern classification view [2], searching for a good predictive model can be attained following two different paths. The most common path is to employ various multivariate data analysis techniques (e.g., neural network, decision tree, support vector machine), and to compare them by assessing their model performance (e.g., off-training set accuracy) measured via some re-sampling technique, such as 10-fold cross-validation. The most accurate model is then used for prediction. A second path is to keep the multivariate analysis technique fixed and instead work on improving the feature representation, by either selecting the most relevant features (e.g. calorimeter cluster parameters, number of tracks associated with the event, mass of  $\tau$  tracks, etc.), or by constructing new features through a search in the space of possible feature combinations.

Our study follows the second path described above in the identification of  $\tau$  leptons. This step is particularly important because most classification algorithms are highly sensitive to the quality of the feature representation. The presence of irrelevant features and/or features interacting in complex ways demands learning machines with low bias or high capacity (i.e., high flexibility in the decision boundaries), to capture the complex data distribution, at the expense of increasing the variance component in error [3, 5]. By selecting the most relevant features and joining

together highly interacting features the data distribution is transformed to a form amenable to current classification techniques.

Moreover, experiments in particle physics often embed a complex characterization of event signals where evaluating and constructing new features can become highly instrumental. In general, attaining accurate classifiers depends to a great extent on the quality of the feature set characterizing the system under study. On the one side, high quality features convey much information about the system; in this case, a simple classifier suffices to produce good results. In contrast, complex features must be combined with many other features to unveil system structure; here features can interact in many ways and identifying the most relevant features is needed to discover important combinations.

## 2. Multivariate Data Analysis for Classification

We begin by giving a brief overview of the classification problem. We assume an  $n$ -component vector-valued random variable,  $(A_1, A_2, \dots, A_n)$ , where each  $A_i$  represents an attribute or feature; the space of all possible attribute vectors is called the input space  $\mathcal{X}$ . Let  $\{y_1, y_2, \dots, y_k\}$  be the possible classes, categories, or states of nature; the space of all possible classes is called the output space  $\mathcal{Y}$ . A classifier receives as input a set of training examples  $T = \{(\mathbf{x}, y)\}$ , where  $\mathbf{x} = (a_1, a_2, \dots, a_n)$  is a vector or point in the input space and  $y$  is a point in the output space. We assume  $T$  consists of independently and identically distributed (i.i.d.) examples obtained according to a fixed but unknown joint probability distribution  $\phi$  in the input-output space  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ . The outcome of the classifier is a function  $h$  (or hypothesis) mapping the input space to the output space,  $h : \mathcal{X} \rightarrow \mathcal{Y}$ . Function  $h$  can then be used to predict the class of previously unseen

	A	A'	
S	$n_1^1$	$n_1^0$	$n_1$
S'	$n_0^1$	$n_0^0$	$n_0$
	$n^1$	$n^0$	$N$

Probabilities:

$$\text{For } S \text{ and } S': P^1 = \frac{n^1}{N} \quad P^0 = \frac{n^0}{N}$$

$$\text{For } A: P_1^1 = \frac{n_1^1}{n_1} \quad P_1^0 = \frac{n_1^0}{n_1} \quad P_1 = \frac{n_1}{N}$$

$$\text{For } A': P_0^1 = \frac{n_0^1}{n_0} \quad P_0^0 = \frac{n_0^0}{n_0} \quad P_0 = \frac{n_0}{N}$$

Figure 1: Cross-classification of feature values and classes with probabilities estimated from the data.

attribute vectors.

In our study each feature vector  $\mathbf{x}$  stands for a jet of energy flow, characterized by features such as calorimeter cluster parameters, number of tracks associated with the event, mass of  $\tau$  tracks, etc. The class variable is binary-valued: events either belong to a  $\tau$  signal or a background event. Our problem formulation can be exploited by any classification algorithm including neural networks, decision trees, support-vector machines, etc. Rather than focusing on the classification algorithm, however, we focus on the feature representation as described next.

### 3. Feature Evaluation

We first address the problem of selecting the most relevant features. To assess the value of potentially useful features we need an evaluation metric. Formally, an evaluation metric  $M$  is used to quantify the quality of the partitions induced by a feature  $A$  over a training set  $T$ , where  $|T| = N$ . For simplicity assume feature  $A$  is binary-valued, such that it divides  $T$  in two sets:  $\{(\mathbf{x}, y) \mid A(\mathbf{x}) = 1\}$  and  $\{(\mathbf{x}, y) \mid A(\mathbf{x}) = 0\}$ ; we say the former set is covered by  $A$ , whereas the latter set is covered by the complement  $A'$ . Similarly, set  $T$  can be divided according to the class label on each example; we assume two class values: 1 (for signal  $S$ ) and 0 (for background  $S'$ ). Figure 1 shows the cross-classification of classes and values of  $A$ . Let  $n^1$  and  $n^0$  be the number of examples in  $T$  of class 1 and 0 respectively, where  $n^1 + n^0 = N$ . Let  $n_1^1$  and  $n_1^0$  be the number of examples covered by  $A$  of class 1 and 0 respectively, such that  $n_1^1 + n_1^0 = n_1$ , and let  $n_0^1$  and  $n_0^0$  represent the corresponding numbers in  $A'$ , such that  $n_0^1 + n_0^0 = n_0$ . In addition, Figure 1 defines probabilities as estimated from the data.

The quality of the partitions made by  $A$  is simply determined by the class-uniformity or purity of such partitions. Good attributes tend to split dataset  $T$  into subsets that are class-uniform. The following are traditional definitions of evaluation metrics where the goal is to maximize the output value.

#### Information Gain (IG) [11]

Let entropy  $H(x, y) = -x \log_2(x) - y \log_2(y)$

$$\text{IG}(A) = H(P^1, P^0) - \sum_{i=0}^1 (P_i H(P_i^1, P_i^0)) \quad (1)$$

#### Gain Ratio (GR) [11]

$$\text{GR}(A) = \frac{\text{IG}(A)}{H(P_0, P_1)} \quad (2)$$

#### G Statistic (G) [13]

$$G(A) = 2N \text{IG}(A) \log_2 2 \quad (3)$$

#### $\chi^2$ [13]

$$\chi^2(A) = \frac{N (n_1^1 n_0^0 - n_1^0 n_0^1)^2}{n^1 n^0 n_1 n_0} \quad (4)$$

Traditional evaluation metrics as the ones described above define the degree of class-uniformity in each new partition using the proportion of classes on the example subsets; the best result is attained if each example subset is class uniform.

To gain a better understanding of the nature of evaluation metrics, note each metric  $M$  is a function of the number of examples covered by feature  $A$  and of its complement  $A'$ ,  $M : f(n_1^1, n_1^0, n_0^1, n_0^0)$  (Figure 1). Alternatively  $M$  could be defined as a function of the coverage of  $A$  and of the coverage of the whole set  $T$ ,  $f(n_1^1, n_1^0, n^1, n^0)$ , since  $n^1 = n_1^1 + n_1^0$  and  $n^0 = n_0^1 + n_0^0$ . For a given learning problem,  $n^1$  and  $n^0$  are fixed; by considering them as constants, we can simply express  $M$  as  $f(n_1^1, n_1^0)$ . For simplicity let's rename  $n_1^1$  and  $n_1^0$  as  $p$  and  $n$  (the positive –or signal– and negative –or background– examples covered by  $A$ ), such that  $M : f(p, n)$ . Metric  $M$  extends above the plane defined by these two variables. We define this plane as the *coverage plane*. Each of the metrics defined above can be plotted above a coverage plane bounded by the total examples of class 1 and class 0 in  $T$ ,  $n^1$  and  $n^0$ . As an example, Figure 2 plots Information Gain when the number of positive and negative examples in  $T$  is the same ( $n^1 = n^0 = 100$ ); each point  $(p, n)$  is evaluated according to Equation 1. The fact that the value of  $f(p, n)$  takes into account both the coverage of  $A$

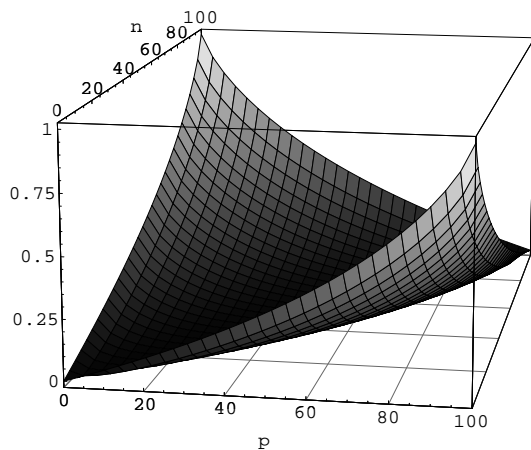


Figure 2: Information Gain as a function of the possible coverage (number of examples of class 1 and class 0) of a feature.

and of its complement  $A'$  is reflected by the symmetry about the axis-line  $((0, 0), (100, 100))$ . The maximum values are attained at the extreme points  $(100, 0)$  and  $(0, 100)$ , when the induced example subsets are class uniform [12].

Although our focus here is on traditional metrics, the reader should be aware of additional families of metrics that quantify the ability of a feature to separate away examples of different class [6, 7, 9]. These metrics deserve particular attention because of their ability to address the high interaction problem, in which the relevance of a feature can be observed only in combination with other features.

### 3.1. Empirical Results

Our first set of experiments rank all available features using information gain (Equation 1) as the evaluation metric. We use Monte-Carlo simulated data with a skewed class distribution (about 5% of events belong to background and the rest to signal). Each feature is first discretized into intervals [1]. Out of twenty one features, each considered separately, only six produce information gain above 0.1. Figure 3 shows histograms corresponding to the posterior class distributions conditioned on the best two features. The first feature, *tautz* (Fig. 3, top), is the coordinate value in the direction of the beam ( $z$ -coordinate) of the most energetic daughter particle produced from a  $\tau$  lepton decay at its closest approach to the  $z$ -axis. The second feature, *tauiso* (Fig. 3, bottom), is the normalized transverse component of  $\tau$  lepton momentum measured based on the distribution of electromagnetic calorimeter cells. Though some overlap is present, both features effectively discriminate between both classes.

Reducing the number of features is not only useful

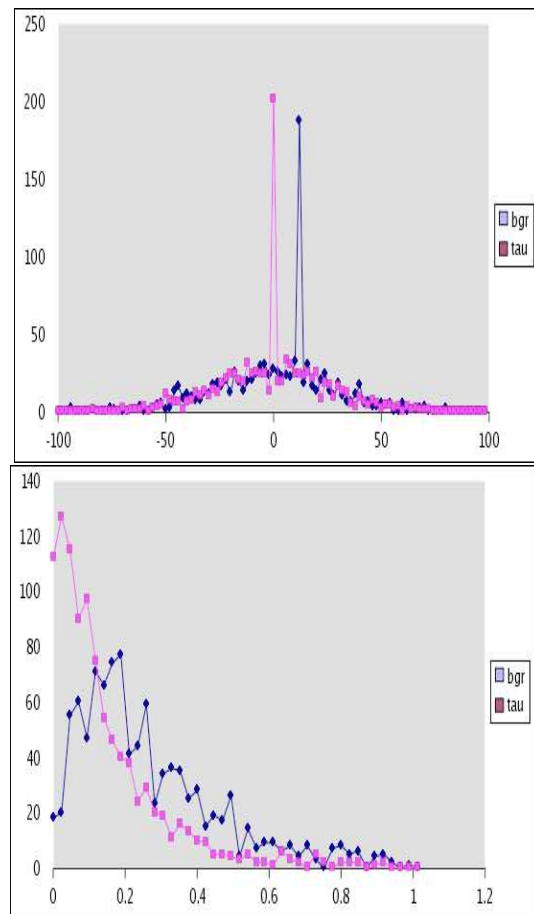


Figure 3: Posterior class distributions conditioned on the two best features, *tautz* (top) and *tauiso* (bottom).

to speed up the training phase, but also eliminates noise. Additional experiments compare the accuracy of several classifiers using the top six features; our results show no significant difference in accuracy between these results and those obtained using all available features. We conclude feature evaluation can be very useful in identifying relevant features when discriminating signal events from background events in particle physics.

## 4. Feature Construction

A second approach to improve the data representation is to construct new features that could potentially capture important dependencies among original features. To begin, let us assume our problem characterization is made of Boolean features. This is attained by dividing numeric features into intervals [1], and by mapping each nominal value into a Boolean feature. One approach to feature construction is to build new

features as the logical combination of Boolean features or their complements (e.g.,  $\bar{a}_1 \wedge a_2$ ).

Our algorithm conducts a search over the space of all logical combinations of features. Each combination can be evaluated using any of the metrics described in Section 3. Since the size of the search space is frequently computationally intractable, we used genetic algorithms to find the highest-ranked feature combinations.

Genetic algorithms operate in a simple fashion: they work iteratively on a population of individuals or candidate solutions with highest fitness value. At each iteration individuals are ranked based on their fitness value (in our case best score output by the evaluation metric, e.g., information gain) and a new population is produced by applying genetic operators on individuals selected probabilistically. The process continues until an individual is found with a fitness value above a predefined threshold. Genetic algorithms are search mechanisms amenable to parallelization, effective in finding solutions in complex spaces [4, 10].

## 4.1. Empirical Results

Our second set of experiments use genetic algorithms to explore the space of logical feature combinations. We observe best results when using Gain Ratio as the evaluation metric (Equation 2). To assess the utility of our results we compare the accuracy of a decision tree with and without the new constructed features. We observe a significant difference in accuracy between the two resulting hypotheses, supporting our claim that constructing new features can improve the original data representation by capturing dependencies among attributes.

## 5. Conclusions

Choosing the right feature representation can influence the quality of the predictive model with equal or higher impact than choosing the right classification algorithm. This is because classification algorithms are highly sensitive to the quality of the feature representation. In this paper we show how feature evaluation can be used to systematically rank features according to their correlation with the class under prediction (i.e., their correlation with  $\tau$  lepton signals).

Our experiments use Monte-Carlo simulated data with two types of events:  $\tau$  lepton signals and background events. Our evaluation process shows how only a few features are necessary to produce a classifier. Such reduction in the size of the feature set reduces the time to train the classifier and often results on an improvement in the accuracy of the final hypothesis.

We also show how to automatically construct new features by exploring the space of logical feature

combinations using genetic algorithms. Our results show an improvement in predictive accuracy when the newly constructed features are integrated into the pool of original features. Although feature construction can become computationally expensive, the resulting combinations may point to interesting relations about the physical processes involved in particle collisions and decay.

## Acknowledgments

We thank the DØ collaboration for sharing the Monte-Carlo simulated data used in our experiments.

## References

- [1] J. Catlett (1991). "On Changing Continuous Attributes Into Ordered Discrete Attributes", *Proceedings of the European Conference on Machine Learning* pp. 164-178. Springer-Verlag.
- [2] R. O. Duda, P. E. Hart, and D. G. Stork (2001). "Pattern Classification", John Wiley Ed. 2nd Edition.
- [3] S. Geman, E. Bienenstock, and R. Doursat (1992). "Neural Networks and the Bias-Variance Dilemma", *Neural Computation*, 4, pp. 1-58.
- [4] D. Goldberg (1989). "Genetic Algorithms in Search, Optimization, and Machine Learning", Addison Wesley.
- [5] T. Hastie, R. Tibshirani, and J. Friedman (2001). "The Elements of Statistical Learning, Data Mining, Inference, and Prediction", Springer-Verlag.
- [6] S. J. Hong (1997). "Use of Contextual Information for Feature Ranking and Discretization", *IEEE Transactions of Knowledge and Data Engineering*.
- [7] K. Kira and L. Rendell (1992). "A Practical Approach to Feature Selection", *Ninth International Workshop on Machine Learning*, pp. 249-256, Morgan Kaufmann.
- [8] B. Knuteson and P. Padley (2003). "Statistical Challenges with Massive Data Sets in Particle Physics", *Unpublished Manuscript*.
- [9] I. Kononenko and S.J. Hong (1997). "Attribute Selection for Modeling", *Future Generation Computer Systems*.
- [10] T. Mitchell (1997). "Machine Learning", McGraw-Hill.
- [11] J. R. Quinlan (1994). "Programs for Machine Learning", Morgan Kaufmann, San Francisco.
- [12] R. Vilalta and D. Oblinger (2000). "A Quantification of Distance-Bias Between Evaluation Metrics in Classification", *Proceedings of the 17th International Conference on Machine Learning*, pp. 1087-1094, Morgan Kaufman.
- [13] A.P. White and W. Z. Liu (1994). "Bias in Information-Based Measures in Decision Tree Induction", *Machine Learning*, 15, pp. 321-329, Kluwer, Boston, MA.