PROCEEDINGS OF SCIENCE

PARADIGM, a Decision Making Framework for Variable Selection and Reduction in High Energy Physics

Sergei V. Gleyzer* **and Harrison B. Prosper**
Florida State University, Tallahassee, Florida, United States
E-mail: sergei.gleyzer@cern.ch, harry@hep.fsu.edu

In high energy physics, variable selection and reduction are key to conducting robust multivariate analyses. Initial variable selection often results in variable sets with greater cardinality than the number of degrees of freedom of the underlying model. This motivates the need for variable reduction, and more fundamentally, for a consistent decision making framework. Such a framework called PARADIGM, based on a global reduction measure called the global loss function and relevant for searches for new phenomena in physics, is described in detail. We illustrate the common pitfalls of variable selection and reduction, such as variable interactions and variable shadowing, and show that PARADIGM gives consistent results in their presence. In this paper, we discuss the application of PARADIGM to several searches for new phenomena in high energy physics and compare the performance of different measures of relative variable importance, in particular of those based on binary regression. Finally, we describe a technique called variable amplification and show how PARADIGM can be used to improve classification performance.

XII Advanced Computing and Analysis Techniques in Physics Research
November 3-7 2008
Erice, Italy

*Speaker.

PoS(ACAT08)067

1. Introduction

Large variable sets are a common occurrence in modern scientific research[1]. Several questions are often encountered that require resolution. Are all the features necessary to achieve a particular analysis performance goal? If a variable set were to be reduced, what is the optimal size of the final feature set? Is there any tolerance to noise? What is the optimal analysis strategy? Some questions are easier to address than others but, fundamentally, a consistent decision making framework is highly beneficial for such circumstances.

In this paper, we propose a decision making framework called PARADIGM, aimed at feature selection, reduction and the improvement of the classification process. PARADIGM provides the researcher with easy to interpret criteria for making decisions relevant to different analysis tasks and features selected for the tasks. The decision making framework is not limited to problems encountered in high energy physics (HEP) but has been developed on the basis of several HEP analyses of varying complexity.

PARADIGM relies on several concepts that have their roots in information and decision theories. First is called *relative variable importance*, useful for tasks not associated with parameter space reduction and also used in PARADIGM in the variable amplification algorithm. The other is the *global loss function*, relevant for parameter space reduction and classifier selection.

In Section 2 we describe the initial classifier selection. Relative variable importance is described in detail in Section 3.1. Sections 3.2 and 3.3 are devoted to the classification process using variable amplification, a novel way of boosting. In Section 3.4 we discuss the challenges frequently encountered in multivariate analysis, such as variable interactions and variable shadowing, together with how PARADIGM helps address them with relative ease. The global loss function is described in detail in Section 3.5, while Section 3.6 discusses optimal classifier selection with the global loss function. Finally, the full decision making framework is summarized in Section 4.

2. Initial Classifier Selection

Classification-based criteria are widely used for variable selection and related decision making [1, 2, 3]. Other measures can be derived directly from data without the use of classification, as in [4]. The classification-independent approach is well known to be robust but less accurate than its classification-based counterpart [1].

PARADIGM is by design classifier-choice independent. A researcher can and should initially choose any or all of the classifiers available to her, such as neural networks, decision trees or rule ensembles, as long as a performance measure can be assigned to all or some of the classifiers. A common choice for this performance measure is the area under the receiver operating characteristic (ROC) curve [5], but other measures may be better [6]. As will be described in Section 3.6, PARADIGM allows the researcher to unambiguously choose the optimal classifier based on the *global loss function* results.

[1] An interesting subject in itself, not discussed in this paper

3. Formulation

3.1 Relative Variable Importance

Relative variable importance reflects the relevance of a particular variable to a given task relative to all other variables. PARADIGM's relative variable importance exhibits the main virtues of other relative importance algorithms [3, 4, 7] such as linear separability and order-independence, and provides additional sensitivity from the inclusion of individual variable effects in classification and the capability to identify noisy and adverse features.

For the initially chosen variable set $\{V\} = \{X_1, ..., X_N\}$, *relative variable importance* (RVI) is defined to be:

$$RVI(X_i) \equiv \sum_{S \subseteq V : X_i \in S} F(S) * W_{X_i}(S) \ , \tag{3.1}$$

where $F(X_i, ...X_j)$ is a general classifier performance measure[2], the sum encompasses subsets $\{S\}$ of $\{V\}$ that contain the variable X_i, and

$$W_{X_i}(S) \equiv 1 - \frac{F(S - \{X_i\})}{F(S)} \ , \tag{3.2}$$

is a weight that accounts for individual variable's share of the classifier performance measure $F(\mathrm{S})$. This weight is defined as a fractional performance loss (or gain) in $F(\mathrm{S})$ if the variable X_i is removed from a classifier. The final RVI values are additionally normalized:

$$N \equiv \sum_{X_i} F(S) * W_{X_i}(S) \ . \tag{3.3}$$

so that all the RVI sum to 1.

A technique with a similar goal in mind is found in the context of rule-based regression, a framework that condenses classifiers (usually decision trees) into sets of (*if, then, else*) rules that can be used as ensemble predictors [3]. Notably, even if all the rules are poor but marginally better than random guessing, in a large ensemble they become a very good predictor [3, 7, 8]. A binary rule-based regression tool called RULEFIT [3] is selected for a comparative study.

3.2 Comparison between RVI and RULEFIT

Analogous to how relative variable importance is defined in RULEFIT [3], the RVI is directly tied to the performance of classifiers containing the variable in question. However, in contrast to RULEFIT, the weight W_{X_i} allows the RVI to be more sensitive to the effects of individual variables during classification and permits the identification of features that have a negative effect on classification.

A typical plot of the RVI for the 27 variables of a representative high energy physics analysis [9] is shown in Fig. 1. On an absolute scale, PARADIGM's RVI exhibits both similarities and differences to RULEFIT's variable importance measure (Fig. 2). Overall, the two criteria appear consistent with one another. The notable exception is a variable on the extreme right of Fig. 2,

[2]The range of the performance measure may vary. For the area under the ROC curve the range of $F(X_i, ...X_j)$ is from 0.5 to 1

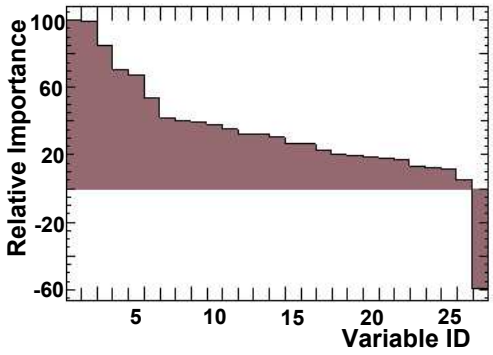

Figure 1: A typical plot of relative variable importance. Variable ID is assigned in the decreasing order of RVI

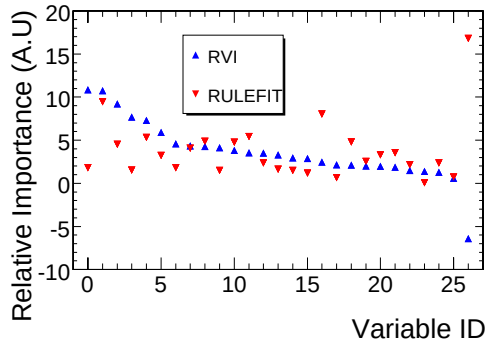

Figure 2: Comparison between variable importance measures provided by PARADIGM and RULEFIT on an absolute scale

unambiguously identified by PARADIGM to be adverse to classification. As is true for all absolute value criteria, the sign of variable importance is unattainable with RULEFIT.

In order to gain insight into relative variable importance, it is worthwhile to consider what makes one relative variable importance measure preferable to another. Comparisons, such as that in Figure 2, can be used to infer the differences and similarities between the two candidate measures, but nothing more.

3.3 Relative Variable Boosting

The optimal way to address this quantitatively rather than qualitatively is to consider the amount of useful information provided by the two criteria and show how that information can be used to achieve analysis goals, for instance, to maximize the classification power of a given set of variables.

We proceed to feed back the relative variable importance information into the classification process, a procedure called RVBoost, that stands for relative variable importance boost or ampli-

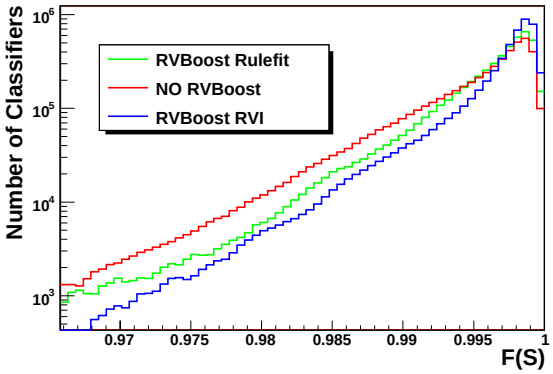

Figure 3: Comparison of classifier performance using RVBoost with PARADIGM and RULEFIT to that without RVBoost, for a fixed number of classifiers

fication. This approach requires creation of new classifiers that use relative variable importance in direct decision making during the classification process. For example in decision trees relative variable importance information is introduced at each decision-making junction, to influence the votes used to split the branches. The same boosting procedure can be easily applied to all known classifiers.

If the relative variable importance measure in question contains information that may be used to further the classification goals of the analysis, it is clearly beneficial. That is the case for both relative variable importance and RULEFIT but in differing amounts (Fig. 3). A fixed number of classifiers in this example is RVBoosted with the RVI and the RULEFIT measures. The performance of the new classifiers is compared to that of the original classifiers. As Figure 3 shows, PARADIGM's RVI criteria outperforms RULEFIT's variable importance and both outperform the original classifiers when relative variable importance boosting is considered.

3.4 Subtlety in Variable Reduction

To illustrate a common caveat in multivariate analysis involving classification, the last variable in Figure 4a is removed and the figure itself redrawn in 4b. A variable that was previously marginally useful became adverse to classification and the order and magnitudes of relative variable importance have changed. This behavior can be explained by the presence of multiple interactions among variables, a common behavior during classification. For instance, in decision trees this can lead to a phenomenon known as variable shadowing, when a presence of one strongly interacting variable partially or entirely shadows its interacting partner, making it appear irrelevant.

The fact that interacting variables influence the performance of their partners in both directions, can be used explain the common occurrence illustrated in Fig. 4. A classical formulation of variable interactions on the basis of risk analysis is found in Ref. [10]. There are several methods to quantify the strength of variable interactions. For example, RULEFIT uses the concept of partial dependencies [3].

As Figure 4 shows, the variable importance landscape becomes distorted by the removal of interacting variables. Presence of variable interactions significantly reduces the effectiveness of

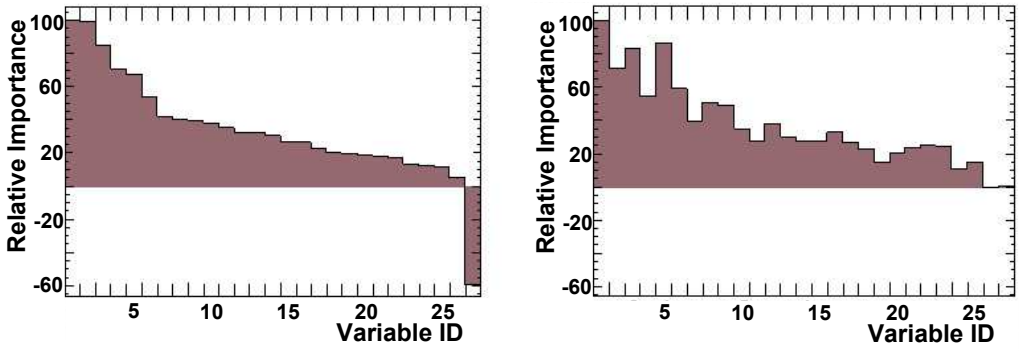

Figure 4: The RVI landscape a) prior to removal of the final variable on the right b) after its removal

criteria that do not directly take them into account, such as RULEFIT's variable importance or the RVI, when it comes to parameter space reduction. Ignoring this subtlety is a common mistake researchers make. One instead should choose measures for parameter space reduction that implicitly incorporate variable interactions, such as the *global loss function*, described in the next section.

3.5 Global Loss Function

The *global loss* or *gloss function* (GF) is an information measure specific to variable reduction that allows a researcher to make sound decisions by incorporating variable interactions. Given a subset to be reduced, the *global loss function* measures the predictive power loss relative to an upper bound of achievable performance of classifiers that remain:

$$GF(S') \equiv 1 - \frac{\sum_{S \subseteq (V-S')} F(S)}{2^{|V-S'|}} \ , \tag{3.4}$$

where $S' \subset V$ is the subset considered for reduction and the absolute scale limit in the denominator is given by[3]:

$$\sum_{S \subseteq V-S'} F(S)_{max} = 2^{|V-S'|}. \tag{3.5}$$

In other words, given the initial variable set {V} from which a variety of classifiers can be built, how much classification performance would be lost if one removes subsets $\{S'\}$ of {V} of various sizes? The answer is precisely the *gloss function*. The lower its value, the lower the loss of classification power resulting from the removal of the subset $\{S'\}$.

A characteristic plot of the global loss function is shown in Fig. 5. Note, that in this figure the $\{S'\}$ subsets are ordered by increasing cardinality, and within the regions of equicardinality, they are arranged by a binary index[4]. Therefore, the subset $\{S'\}$ on the extreme left of Fig. 5a is the null set $\{\emptyset\}$, for which the GF is approaching 0 but is still finite. On the opposite extreme is the set $\{V\}$ with the maximum *gloss function* value of 0.5, which reflects the fact that all the variables have been removed.

[3]that follows from: $\sum_{k=0}^{n} \binom{n}{k} = 2^n$ and $F(S)_{max} = 1$. If $F(S)_{max} \neq 1$, the right hand side of Eq. 3.5 and the denominator in Eq. 3.4 instead become $2^{|V-S'|} \times F(S)_{max}$

[4]index, where each digit signifies the presence or absence of a corresponding variable in $\{S'\}$.

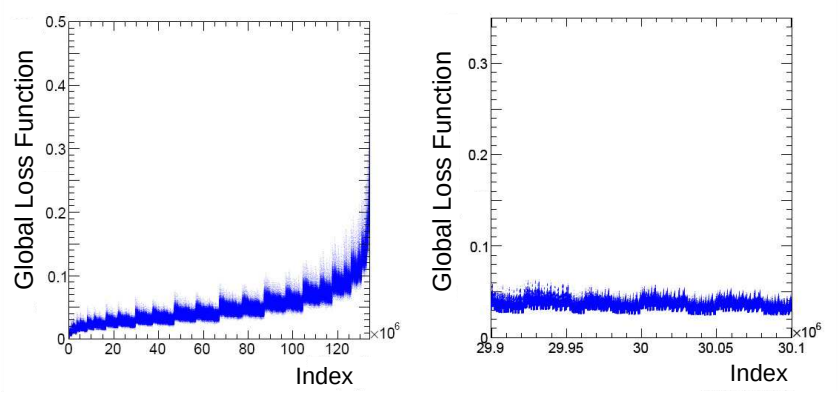

Figure 5: Left: the gloss function for $0 \leq |S'| \leq 27$. Right: a snapshot within the equicardinality region $|S'|=15$

3.6 Optimal Classifier Selection

The *global loss function* permits quantitative selection of optimal classifiers for given analysis tasks, such as, for example, searches for previously unseen phenomena. The lower the area under the gloss function curve (Fig. 5), the better the classifier choice this task. Different classifier choices can be easily compared on this absolute scale leading to an optimal choice for further analysis.

4. Decision Making Framework

By combining the *gloss function* and *relative variable importance*, a powerful decision making framework can be made. The structure of the framework is as follows:

- As described in Section 2, suitable classifiers are selected
- An optimal classifier is chosen with the *global loss function*
- If parameter reduction is desired, the $\{S'\}$ subset with the minimum gloss function value is chosen for reduction and its compliment is kept for further analysis
- Relative variable importance is used to cross-check that all adverse variables are included in the $\{S'\}$ subset to be reduced
- Once the final variable set is selected, the RVBOOST procedure described in Section 3.3 is applied to maximize the performance of the classifiers built from this set

5. Discussion and Summary

It is worthwhile to note that minimization of the *global loss function* is not equivalent to maximization of the classifier performance measure F(S), i.e. finding the highest performing classifier and its constituent variables (Fig. 6). Some researchers attempt a quick search for high performing

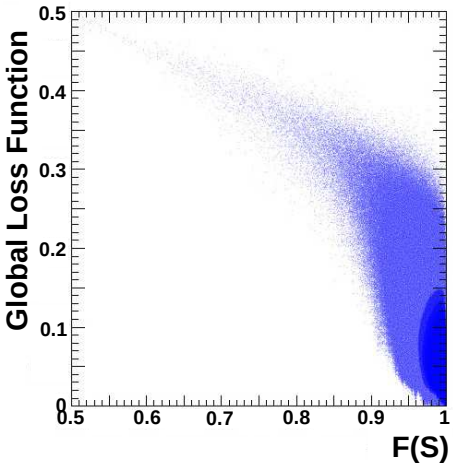

Figure 6: The non-linear relationship between the global loss function and the classifier performance measure $F(S)$

classifiers, typically by adding or subtracting variables with forward selection/backward elimination methods [11]. Once such a classifier is found its constituent parameter space is declared optimal for further analysis. This approach, besides neglecting the variable interactions, is inflexible.

In realistic searches for new phenomena that occur in nature classifiers are typically trained on simulated (usually Monte-Carlo) data for at least one of the major classes of events, usually the one related to the previously unseen object or model. If the researcher limits herself to only one classifier, or alternatively to only one of the many possible combinations of the reduced parameter space, without considering the associated loss of information, she becomes limited in options if the search does not yield the desired result. Remaining options are to set limits and start over. Making a choice of the parameter space based on the global loss function criterion, that consistently produces a strong family of classifiers out of its constituents, allows one to step back and modify the parameter space slightly and maintain a required high performance level, without having to repeat the classifier search. This becomes crucial when the models that are being probed come in significant variety and contain free parameters with unknown values. In this case, flexibility, tied with high performance, becomes a crucial aspect of a successful search.

In summary, PARADIGM is a robust parallelized framework that provides decision-making information to assist and improve modern day multivariate analyses. PARADIGM 2.0 is the software version used for this study. Its areas of application are classifier selection, classifier improvement, variable selection and variable reduction. As the next step, the authors (in particular S.G.) would like to implement or help implement the algorithm in a multivariate analysis framework.

References

[1] A.L. Blum, P. Langley, *Artificial Intelligence* **97**, 245 (1997).

[2] R. Kohavi, G.H. John, *The Wrapper Approach* in Feature Selection for Knowledge Discovery and Data Mining, edited by H. Liu & H. Motoda (Kluwer Academic Publishers, Norwell MA) (1998).

[3] J.H. Friedman, B.E. Popescu, Technical Report, Statistics Department, Stanford University. For details, see http://www.stat.stanford.edu/people/faculty/friedman/index.html (2005).

[4] W. Kruskal, *The American Statistician* **41**, 6 (1987).

[5] J. Bradley, *Pattern Recognition* **30**, 1145 (1997).

[6] J. Makhoul, F. Kubala, R. Schwartz, R. Weischedel, in proceedings of the *DARPA Workshop on Broadcast News Understanding*, 37 (1999).

[7] L. Breiman, *Machine Learning* **24**, 123 (1996).

[8] L. Breiman, *Machine Learning* **45**, 5 (2001).

[9] V.M. Abazov et al., *Phys. Rev. D* **78**, 012005 (2008).

[10] L.A. Cox, *Management Science* **31**, 800 (1985).

[11] R. Kohavi, in proceedings of *First International Conference of Knowledge Discovery and Data Mining*, 192 (1995).