A new method for identifying neutrino events in IceCube data

Dmitry Chirkin[*]

[*]*University of Wisconsin, Madison, U.S.A.*

***Abstract*. A novel approach for selecting high-quality muon neutrino events in IceCube data is presented. The rate of air shower events mis-reconstructed as signal is first reduced via the use of the geometrical (software) trigger. The final event selection is performed with a machine-learning method, designed specifically for IceCube data. It takes into account some generic properties of IceCube events, e.g., the fact that separation of signal from background is more difficult (requiring tighter cuts on the quality parameters) for horizontal rather than vertically up-going tracks. The method compares favorably to other techniques in situations with both high and low simulation statistics.**

***Keywords*: neutrino search, machine learning, event selection**

I. INTRODUCTION

An important task of a neutrino telescope like IceCube is identifying extra-terrestrial neutrinos that are interspersed between orders of magnitude higher background of particles originating in the showers produced by cosmic rays in the Earth's atmosphere.

As a first step a high purity atmospheric (plus possible extraterrestrial) neutrino event sample is selected, with only a small contaminating fraction of mis-reconstructed atmospheric muon events. Only neutrinos can cross the overburden of the Earth in the upward direction; however, selecting events reconstructed as upward-moving leaves many mis-reconstructed atmospheric muons in the sample, improving the ratio of neutrino to contaminating muon events (initially at $\sim 10^{-6}$) by only a factor of ~ 100.

The problem is further exacerbated by a highly uneven contamination of the mis-reconstructed muons in several of the analysis variables, most importantly the zenith angle. This contamination is smaller for up-going directions and increases for more horizontal tracks, growing rapidly near and above the horizon. It is therefore difficult to arrive at an event selection method that provides optimized cut surfaces simultaneously for all zenith angles. Splitting the cut optimization in different zenith bins leads to fluctuations of the cut parameters from one zenith angle bin to the next that are perceived as unphysical. In situations with limited simulated data, splitting it in several zenith angle bins is undesirable.

This author has also performed an SVM-based event selection optimization and found that training the SVM gets more difficult for zenith angle ranges extending above the horizon.

The above considerations led to the development of a new framework for selecting and applying cuts on quality parameters, that in the following is called "Subset Browsing Method", or *SBM* for short.

The quality parameters used with the event selection method of this paper build upon those discussed previously [1].

II. SIMPLE EXAMPLE

First consider a simple example employing only two parameters that select events with lower background contamination for lower value of the quality parameter. These can be, e.g., zenith angle (0 degrees for up-going to 90 for horizontal tracks) and estimated angular resolution (e.g., describing the half-width of the likelihood function at the minimum corresponding to the reconstructed track direction). Both of these can be used to remove the background of mis-reconstructed events, one through the basic reconstruction property, and the other though our prior knowledge that the contamination is higher for tracks near the horizon.

The toy simulated events are divided into two groups (randomly): the training set that is used for the training of the machine, and the testing set, that is used to judge its performance. Both sets, while drawn from the same parent distribution, are statistically independent. The toy "data" events are also simulated and drawn from a somewhat wider signal distribution (to demonstrate the effect of cuts in the transitional region between signal and background). The 3 steps of the machine application are the steps 1^a, 1^b, and 2 as shown on Figure 1.

The training of the machine in this simple example is achieved by identifying the "outlying" background events (on the signal side of the distribution), and creating the "angle cuts" (shown with black straight lines) by cutting away everything on the rejected sides of such cuts (i.e., everything above and to the right of the background event, including that background event itself).

The cuts so identified will obviously remove all background events in the training dataset. As seen from the second row of Figure 1, these cuts do not remove all of the background events when applied to the testing dataset, so a further step, here called 1^b is necessary. Using the angle cuts derived in step 1^a a quality parameter (*SBM**) is constructed, which is simply the count of "angle cuts" of step 1^a that fall into the bad quadrant (up and to the right) of a tested event, see Figure 2. This quality parameter could also be constructed as a

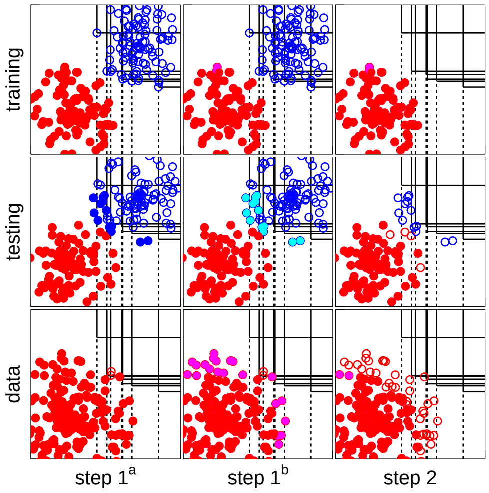

Fig. 1. Two event populations are shown: red is signal for the training and simulated testing event sets, and all data in the data set. Blue points (located up and to the right from red points) describe background events. Lower values on both x and y mean better quality. In steps 1^a and 1^b empty circles show background events removed by the *skeleton cuts*, aqua and pink points show background and signal (or data) events respectively that are removed by the machine quality parameter cut set at 1.5. In the third column events removed by step 2 are shown as empty circles and events additionally removed by the quality parameter cut are shown in pink, same as before. The black lines show the *skeleton cuts* and go through the out-most background events of the training set.

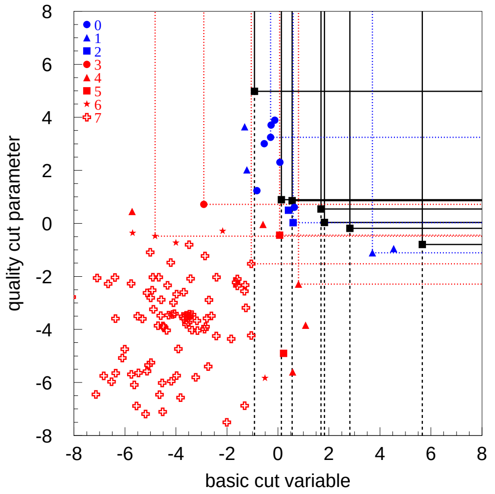

Fig. 2. Shown are events remaining in the testing dataset after the application of step 1^a, signal in red, and background in blue. The legend indicates the value of the *SBM** quality parameter for the shown events. For each value of the *SBM** quality parameter the bad sides of a representative event are shown with straight lines. The *SBM** quality parameter is simply the count of the "outlying" background events of the training dataset that gave rise to the "angle cuts" of step 1^a (indicated with black squares and black solid lines).

weighted sum, as described in the following section, shown for comparison as *SBM* in Figure 4.

A map of the quality parameter (*SBM**) is shown in Figure 3. It is clear that through application of the quality parameter some space is inserted between the cut structure achieved in step 1^a and events with quality parameter greater than 0. Figure 4 shows that a value of *SBM**=2.5 completely separates signal from background in the testing dataset of this simple example.

III. Unsimulated events and step 2

After the application of steps 1^a and 1^b to the real IceCube data it became obvious that, although much of the background events like those present in our simulation was removed, some "unsimulated" background remained in data. This affected agreement in parameter distributions and was particularly evident upon visual inspection. One class of such events appeared to contain two or more coincident but independent muon hit patterns that happened near each other with much overlap in time and failed to be split by the topological trigger. With more simulation we would most likely have been able to correctly identify these events.

Another class of events appeared to contain a bright electromagnetic or hadron shower (*cascade*), with rate of occurrence higher than that predicted by the simulation. While these events, when understood, could be very interesting, making a valuable contribution to the final event selection, it is unclear at this point whether they should be classified as signal or background. Moreover, since they are not simulated as either, the detector effective area to these events cannot be estimated, so they do not contribute to any of the physics results.

A common way to deal with this is by raising the quality parameter of an analysis past the point where all simulated background is removed and to the point where an agreement between data and signal simulation is achieved. This is possible if the quality parameter judges not only how far a given event is from the background region, but also how close it is to the signal region.

In the method presented here the quality parameter is constructed using only the information about the outlying background events that form the "angle cuts" (after some initial amount of signal events is carved out by the step 1^a). Thus, the quality parameter judges only the distance to the background region (contrary to other approaches). To achieve the "similarity with signal events" another step (called step 2) becomes necessary.

This event selection step is achieved by removing all data events, to the bad sides of which there are no signal events of the training set (remaining after the application of steps 1^a and 1^b). The effect of step 2 is demonstrated in the last column of Figure 1: all data events on the background side of the signal region are removed.

IV. Multi-dimensional generalization

First we re-iterate that the SBM method relies on the important observation that most of the quality parameters used in the analysis of IceCube data have the following property: as the fits become less constrained at lower number of channels N_{ch} or strings N_{str} (that received hits), the cuts on the quality parameters necessary to reach a given signal purity need to be tightened. Alternatively, the cuts applied to quality parameters of events with higher N_{ch} or N_{str} can be relaxed somewhat. A

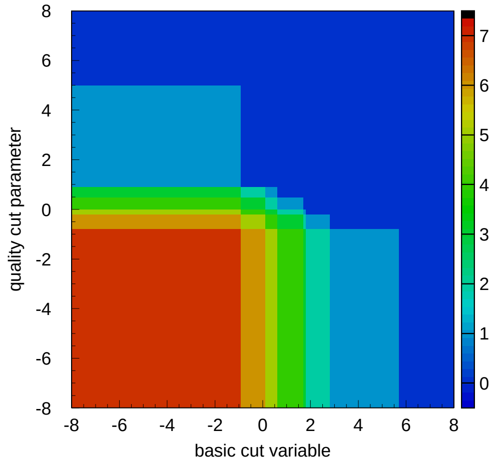

Fig. 3. Map of the quality parameter (*SBM**) calculated according to prescription of Figure 2. Highest-quality region is shown in red.

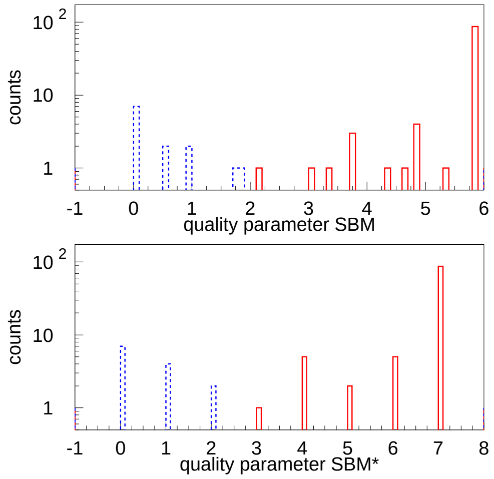

Fig. 4. Quality parameters: simple sum over the "angle cuts" (*SBM**), and weighted sum (*SBM*). Red solid and blue dotted lines show the distribution of signal and background events, respectively.

similar behavior of cuts on quality parameters can be argued for their dependence on the reconstructed zenith angle θ: at angles closer to the horizon the number of background events seeping through is higher than for tracks going up closer to the vertical, so to reach the same purity the cuts on the quality parameters need to be tighter for events with higher reconstructed zenith angle θ. To summarize, we introduce the following

Basic cut property: the cuts necessary to reach the same signal purity satisfy the following conditions:

$$c(\theta^\star, N_{ch}, N_{str}) \le c(\theta^0, N_{ch}, N_{str}) \text{ for } \theta^\star \ge \theta^0$$

$$c(\theta, N^\star_{ch}, N_{str}) \le c(\theta, N^0_{ch}, N_{str}) \text{ for } N^\star_{ch} \le N^0_{ch}$$

$$c(\theta, N_{ch}, N^\star_{str}) \le c(\theta, N_{ch}, N^0_{str}) \text{ for } N^\star_{str} \le N^0_{str}$$

This relies on the **assumption** that lower cut values imply tighter cuts[1]. Parameters θ, N_{ch}, and N_{str} that allow such a behavior of cuts are in the following called *basic cut variables*. The following discussion is simplified with a

Definition: a cut $c^\star$ defined for a set of events with $\theta^\star$, $N^\star_{ch}$, and $N^\star_{str}$ is said to be operating on a *subset* of events of a cut c^0 defined for a set of events with θ^0, N^0_{ch}, and N^0_{str} if $\theta^\star \ge \theta^0$, $N^\star_{ch} \le N^0_{ch}$, and $N^\star_{str} \le N^0_{str}$.

Main cut property: a cut operating on a given set of events also operates on all its subsets.

To rephrase, a cut c^0 defined for a set of events with θ^0, N^0_{ch}, and N^0_{str} also applies to any set of events with $\theta^\star$, $N^\star_{ch}$, and $N^\star_{str}$ (that has its own cut $c^\star$), if $\theta^\star \ge \theta^0$, $N^\star_{ch} \le N^0_{ch}$, and $N^\star_{str} \le N^0_{str}$. To prove we need to show that $c^0 \ge c^\star$. Using 2 intermediate sets of events, and the basic cut property introduced above

$$c^0 = c(\theta^0, N^0_{ch}, N^0_{str}) \ge c(\theta^\star, N^0_{ch}, N^0_{str}) \ge c(\theta^\star, N^\star_{ch}, N^0_{str}) \ge c(\theta^\star, N^\star_{ch}, N^\star_{str}) = c^\star$$

This property allows us to consider all cuts as operating not only on events with $\theta = \theta^0$, $N_{ch} = N^0_{ch}$, and $N_{str} = N^0_{str}$, but rather on all events with $\theta \ge \theta^0$, $N_{ch} \le N^0_{ch}$, and $N_{str} \le N^0_{str}$.

Definition: The cut c^0 associated with θ^0, N^0_{ch}, and N^0_{str} is considered *redundant* if there exists another cut $c^\star$ associated with some other $\theta^\star$, $N^\star_{ch}$, and $N^\star$ such that

$$c^\star \le c^0,\ \theta^\star \le \theta^0,\ N^\star_{ch} \ge N^0_{ch}, \text{ and } N^\star_{str} \ge N^0_{str}.$$

This is because the new cut $c^\star$ clearly implies c^0 by the main cut property.

For each background event i_b in the simulated training dataset its quality parameters are used to create n cuts associated with θ^{i_b}, $N^{i_b}_{ch}$, and $N^{i_b}_{str}$ of the event. The signal purity $p^{i_b} = s^{i_b}/(s^{i_b} + b^{i_b})$ of the events with $\theta \ge \theta^{i_b}$, $N_{ch} \le N^{i_b}_{ch}$, and $N_{str} \le N^{i_b}_{str}$ is then calculated and used to find the i_b that defines cuts in a region with the worst purity. Out of the n cuts associated with i_b the cut that results in a smallest loss of signal events is then chosen and applied to the whole subset on which this cut operates. To accelerate this process if a cut in encountered that removes no signal events it is immediately used without taking into account the purity of the subset of events on which this cut operates.

This procedure is then repeated until the background events in the simulated training dataset are exhausted. At that point all cuts of all background events are cycled through once again, and those that result in no further loss of remaining signal events (which are said to form a *core* or signal events) are saved into the *trained cut set* of the machine. One can further *reduce* this set by removing the *redundant* cuts from it, thus resulting in an *irreducible trained cut set*, which is the result of this machine training procedure.

One can obviously remove all background events (i.e., reach a 100% signal purity) by applying all cuts from the *irreducible trained cut set* to the simulated training dataset. However, when the same is applied to the separately generated *testing* dataset a number of

[1] Some of the quality parameters may need to be taken with a minus sign or as one over their value to satisfy this assumption

background events seep through and the signal purity never reaches 100%.

This may happen, e.g., if we encounter a background event with, say, $N^{\star}_{ch}$ higher than $N^{i_b}_{ch}$ of every background event in the simulated training dataset, thus there are no cuts available that would remove such an event from the testing dataset.

A way around this is to find at least one cut c with $\theta^c \geq \theta^{\star}$, $N^c_{ch} \leq N^{\star}_{ch}$, and $N^c_{str} \leq N^{\star}_{str}$ such that $c^{\star} = q^{\star} \leq c$ ($q^{\star}$ being the quality parameter of the tested event). By the *main cut property* the cut that would achieve the same purity p_c on the subset defined by $\theta^{\star}$, $N^{\star}_{ch}$, and $N^{\star}_{str}$ as the cut c on the subset defined by θ^c, N^c_{ch}, and N^c_{str} is necessarily no more tight as the cut c. That is, applying the cut c on the subset defined by $\theta^{\star}$, $N^{\star}_{ch}$, and $N^{\star}_{str}$ achieves at least the same or higher level of purity as p_c. Now, if an event defined by its quality parameters $c^{\star} = q^{\star}$ but at the *basic cut variables* θ^c, N^c_{ch}, and N^c_{str} of the cut c is passed by the *trained cut set*, cut c is called the *purity cut* defined for the original event (defined by its own $c^{\star}$, $\theta^{\star}$, $N^{\star}_{ch}$, and $N^{\star}_{str}$).

Existence of at least one such cut c for each of the testing dataset events passed by the machine guarantees that the purity in the regions with extrapolated θ, N_{ch}, and N_{str} is at least as good or better than in the regions for which background events existed in the simulated training dataset. This is an important advantage of the discussed method compared to the other machine learning techniques.

Counting all *purity cuts* available for a given testing dataset event provides one with an important machine quality parameter which value is higher for events that are more likely to be signal and lower for events that are more likely to be background. It appears that a cut on this parameter improves the purity in all subsets by equal amount. In order to improve the purity in all subsets to the same final value one may weight the terms in the quality parameter sum with the initial purity of the simulated training dataset on the subsets of the cuts used in the sum (thus leading to the weighted sum definition of SBM, as shown in Figure 4).

We call the machine learning method described here the *subset browsing method* because of the technique in which one has to *browse* through the *subsets* on which the cuts of the *trained cut set* are defined to calculate the quality parameter separating signal from background. The quality parameter itself is called the *SBM quality parameter*: SBM.

The *irreducible trained cut set* forms a “skeleton” of cuts that are applied to the testing dataset achieving the *initial SBM cut level*: $SBM = 0$. The SBM quality parameter is usually normalized so that the highest value of SBM of a background event in a simulated testing dataset is 1 (e.g., in the plot of the SBM in Figure 5).

V. Conclusions and Outlook

We present a new framework for selecting and applying cuts on the quality parameters, here called SBM. It is particularly well-suited for application to IceCube data analysis as it takes into account (both during training and for event classification) some obvious relationships between the quality parameters, which are hardwired into the algorithm.

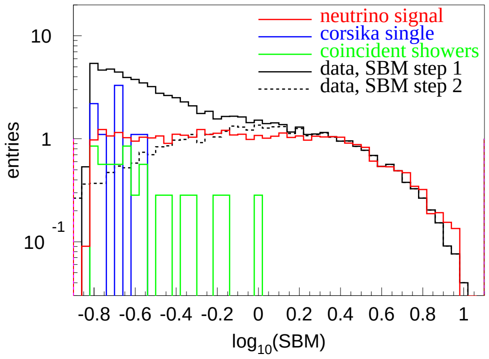

Fig. 5. Distribution of the quality parameter after step 1^a (for 1 year of IC-22 data). The >90% estimated purity is achieved at SBM=0.36.

The SBM appears to separate signal and background well even if the number of simulated training dataset events is low. The SBM extrapolation behavior is very robust and the performance of extrapolation improves as the simulated testing dataset statistics increases, since more *purity cuts* become available to testing events in regions less populated by cuts of the *trained cut set*. The machine will not go around individual background events of the training dataset (a condition that may occur in other methods) because the cuts, by construction, are monotonous functions of the basic parameters.

Additionally, the learning method itself is very simple and has virtually no parameters to set. Most of the machine is implemented with only an application of the $\leq$ operator (and a simple summation for estimating the quality parameter).

The method splits the classification of the data events into two steps: dissimilarity with background, and similarity with signal, thus allowing one to investigate possible “unsimulated” classes of events.

This method was used to identify atmospheric neutrino events in IceCube data taken during the 2007 operation season (see Figure 5 and reference [2]).

As an outlook, this method may be well-suited to analyzes that depend on a veto region around the interesting events in the detector, as most of the cuts on the quality parameters can be relaxed as the veto region is expanded, thus satisfying the basic cut property if veto size is chosen as a basic cut variable.

The method could be further improved in the future by implementing a technique similar to boosting of the BDT.

References

[1] D. Chirkin, et al., *Effect of the improved data acquisition system of IceCube on its neutrino-detection capabilities*, 30th ICRC, Merida, Mexico (arXiv:0711.0353)

[2] D. Chirkin, et al., *Measurement of the atmospheric neutrino energy spectrum with IceCube*, these proceedings