A Machine Learning Approach for Dark-Matter Particle Identification Under Extreme Class Imbalance

Raymond Sutrisno,[1] Ricardo Vilalta,[1] and Andrew Renshaw[1]
[1]*University of Houston, 4800 Calhoun Rd., Houston TX, USA*
rasutrisno,rvilalta,arenshaw@uh.edu

Abstract. The Darkside-50 collaboration is an international experiment conducted at the Laboratori Nazionali del Gran Sasso in Italy, where low-radioactivity liquid argon is used within a dual-phase time projection chamber to detect weakly interacting massive particles (WIMPS), one of the leading candidates for dark matter. The Darkside-50 experiment faces two main data-analysis challenges: extreme class imbalance and large datasets. In this paper we show how machine learning techniques can be employed, even under the presence of samples exhibiting extreme class-imbalance (i.e., extreme signal-to-noise ratio). In our data-analysis study, the ratio of negative or background events to positive or signal events is highly imbalanced by a factor of 10^7. This poses a serious challenge when the objective is to identify a signal that can be easily misclassified as background. We compare several techniques in machine learning that deal with the class imbalance problem: ROUS, SMOTE, and MSMOTE. Experimental results on real data obtained from the Darkside-50 experiment show very high recall values (~ 0.985), with reasonable performance in terms of precision (~ 0.80) and F1-score (~ 0.875).

1. Introduction

Many candidates have been hypothesized to describe the apparent missing matter in the Universe, all of which would fall under the category of dark matter. It is referred as dark matter because it is non-luminous and direct observation using traditional astronomical techniques is not possible. Instead, its presence has been inferred by its gravitational effect on surrounding luminous matter, as well as the footprints it has left within the cosmic microwave background throughout the history of the Universe. Among the leading hypothesized candidates, weakly interacting massive particles (WIMPs) have become a favorite for experimentalists, since their interaction with normal matter can be predicted and searched in ultra-sensitive detectors. Interacting via only the weak force, a WIMP particle would have the potential of elastically scattering off the nucleus of an atom that is contained inside a detector here on Earth, producing what is called a nuclear recoil, and giving an avenue for the direct detection of a new particle that could explain the dark matter puzzle. These nuclear recoils would be detectable inside detectors such as the DarkSide-50 detector (Agnes et al. 2015), currently operating at the Laboratory Nazionali Gran Sasso in Italy. DarkSide-50 is an ultra-low background liquid argon time projection chamber built specially for the detection of a WIMP recoiling off the nucleus of an argon atom in the detector, and has been instrumental in the search for high- and low-mass WIMPs (Agnes et al. 2018).

The generated light signals coming from the nuclear recoil of a WIMP with the liquid argon inside the DarkSide-50 detector are captured by photomultiplier tubes set in two arrays, at the top and bottom of the detector. Signals recorded by the photomultiplier tubes are used to reconstruct interactions. However, even with the ultra-low background levels in the DarkSide-50 detector, the expected rate of electromagnetic interactions inside the detector are quite large compared to the expected rate coming from WIMP nuclear recoils. These electromagnetic interactions result from beta-particles and gamma-rays interacting with the orbital electrons of the argon atoms in the detector, which give off a slightly different response relative to the nuclear recoil from a WIMP, allowing for the possibility to distinguish a WIMP signal from background events. The very low expected rate of WIMP interactions in DarkSide-50 (less than 10^{-2} per year), coupled with the background rate (order of 10^7 per year), creates an extreme class imbalance between the potential WIMP signal and the background within a data set that is quite large. With this in mind, an approach to classifying the data using machine learning has been explored and described next.

2. Machine Learning for Dark-Matter Particle Identification

In supervised learning or classification, we assume the existence of a training set of examples, $T = \{(\mathbf{x_i}, y_i)\}$, where vector $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ is an instance of the input space $\mathcal{X}$, and y is an instance of the output space $\mathcal{Y}$. The output of the learning algorithm is a hypothesis (or function) $f(\mathbf{x})$ mapping the input space to the output space, $f : \mathcal{X} \rightarrow \mathcal{Y}$. In our case, vector $\mathbf{x}$ corresponds to features characterizing events, including the following: the number of pulses detected, the integral of the first pulse (S_1), the integral of the second pulse (S_2), the position where the event was located within the detector (given by $< x_{pos}, y_{pos}, t_{drift} >$), and the ratio of the integral of the first 90 nanoseconds of the first pulse relative to S_1. The class y can be a Neutron-Recoil (NR) event (positive example), or an Electron-Recoil (ER) events (negative example).

The Class Imbalance Problem. A common difficulty when applying supervised learning is the presence of tasks with highly imbalanced class priors (i.e., extreme signal-to-noise ratio). This is clearly the case when searching for dark matter particle interactions, where the ratio of background to signal events is highly imbalanced by a factor of 10^7. The difficulty comes in identifying a signal that can be easily misclassified as background. Several techniques have been proposed to deal with the class imbalance problem (Japkowicz 2000). In general, many solutions rely on basic operations: undersampling the majority class, or oversampling the minority class (sampling is usually done under a uniform random distribution). We describe the techniques used in our experiments based on these basic operations.

Random Oversampling and UnderSampling Technique (ROUS). The first technique simply oversamples the minority class and undersamples the majority class. Sampling is done under a uniform distribution. The procedure continues until we reach a perfectly balanced class distribution. Oversampling the minority class is known to lead to overfitting, but such adverse scenarios can be reversed when the majority class is simultaneously undersampled.

Synthetic Minority Over-Sampling Technique (SMOTE). Rather than directly oversampling the minority class by creating copies of existing minority-class examples, SMOTE generates synthetic examples by creating new instances along the vectors connecting a minority-class example with the k-nearest-neighbors of the same class (Chawla et al. 2002). The SMOTE algorithm is parameterized by the number of nearest neighbors K, and an integer N representing the percentage of newly generated synthetic examples. The rationale is to oversample the minority class by *spreading* the location of new examples on the input space; this helps to reduce model complexity (i.e., to avoid overfitting).

Modified SMOTE (MSMOTE). MSMOTE is a modified version of SMOTE that attempts to be more selective when oversampling. This is done by classifying neighbor examples into three categories: security, border, and noise. Whereas SMOTE would consider all three types, MSMOTE focuses on security examples only. The designation is determined by examining the classes of the k-nearest neighbors. If all k-nearest neighbors share the same class as the minority-class example under analysis, the example is considered of type security. A mixture of classes in the neighborhood of a minority-class example suggests the presence of noise or of examples at the border of minority and majority class regions; discarding these examples is hypothesized to improve performance.

3. Experimental Setting

We report on a set of experiments using real data from the DarkSide-50 detector. We tackle the class-imbalance problem using the techniques described above (ROUS, SMOTE, and MSMOTE). We consider Neutron-Recoil (NR) events as positive examples and Electron-Recoil (ER) events as negative examples. We use Random Forests (Ho 1995) as the core learning algorithm. Random Forests and Decision Tree learners come from the Scikit-Learn Python Machine Learning library. We invoke the Classification and Regression Tree algorithm (CART; Breiman et al. 1984), with Gini as the splitting criterion.

Every result is the output of 30 training and testing runs. Both training and testing samples are obtained using stratified random sampling with ten thousand examples for training and fifteen thousand for testing. The huge amount of initial data ($\sim 2.4 \times 10^6$ ER events and 2.6×10^4 NR events) leads to data processing with high computational cost. Our experiments make use of a computer cluster with $5,704$ CPU cores in 169 compute and 12 GPU nodes; cpu type is Intel Xeon E5-2680v4, with approx. 40TB of disk space and hundreds of GB in memory.

4. Results and Discussion

We use three performance metrics to assess model quality in the detection of Neutron-Recoil events: precision, recall, and F1 score. The metrics are defined as a function of the number of true positive (tp), true negative (tn), false positive (fp), and false negative (fn) predictions. They are commonly used in classification tasks with skewed class distributions. The definitions are as follows:

$$\text{Recall} = \frac{tp}{tp+fn} \quad \text{Precision} = \frac{tp}{tp+fp} \quad \text{F1 score} = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$$

Figure 1 contains plots of mean recall, precision, and F1 score when using the three techniques designed to handle skewed distributions. The x-axis corresponds to the amount of oversampling as a percentage of the number of minority-class examples (e.g., $N = 600$ means six additional examples are created for every original example

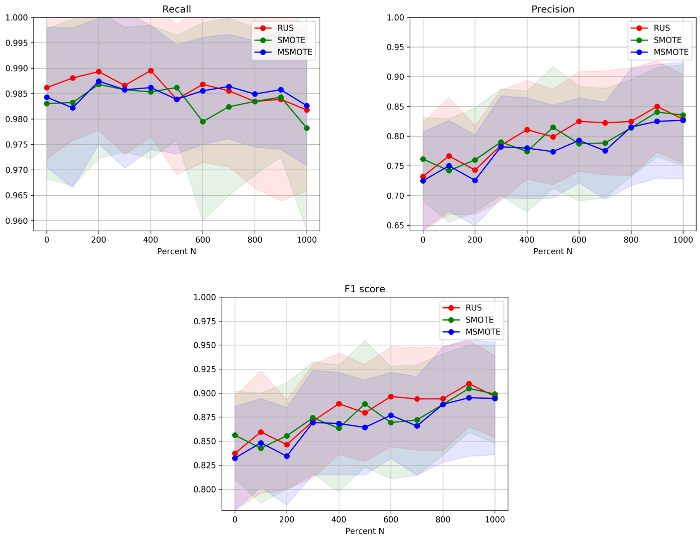

Figure 1. Recall, Precision and F1 score for ROUS, SMOTE, and MSMOTE. Background shadowed regions correspond to +/- one standard deviation estimated from thirty runs.

in the minority class). Results show high values of recall (~ 0.985) that stay relatively constant as oversampling grows; this is indicative of models bearing high sensitivity, where very few signal events are classified as background. Precision, on the other hand, shows lower performance (~ 0.80); it indicates many background events end up classified as signal events. This is expected considering the overwhelmingly small signal to noise ratio. But performance shows improvement as oversampling grows. F1 score (~ 0.875) is simply a harmonic mean of recall and precision and also shows improvement with increased oversampling. In terms of differences across techniques to handle skewed distributions, all of them perform similarly considering the amount of deviation around the mean (shadow regions in Figure 1). We conclude that a simple over- and under-sample technique suffices to handle the class imbalance problem in this particular domain, and that additional work is needed to avoid incorrectly classifying background events as signals.

Acknowledgments. This work was partly supported by the Center for Advanced Computing and Data Systems (CACDS) at the University of Houston.

References

Agnes, P., et al. (DarkSide) 2015, Phys. Lett., B743, 456. 1410.0653

— 2018, Phys. Rev. Lett., 121, 081307. 1802.06994

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. 1984, Wadsworth International Group

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002, Journal of Artificial Intelligence Research, 16, 321

Ho, T. K. 1995, in Document analysis and recognition, 1995., proceedings of the third international conference on (IEEE), vol. 1, 278

Japkowicz, N. 2000, in Proceedinsg of the International Conference on Artificial Intelligence ICAI00