ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

SOM-based Classification and a Novel Stopping Criterion for Astroparticle Applications

Luis Sanchez[1], Erzsébet Merényi[2] and Christopher D. Tunnell[1] *

1- Rice University - Department of Physics and Astronomy

2- Rice University - Departments of Statistics & Electrical and Computer Engineering
6100 Main Street, Houston, Texas - U.S.A.

Abstract. Classification of detector signals is vital in particle physics experiments. However, the intricate spatio-temporal nature of the data and instrumentation effects make highly accurate classification challenging. In this study we use a Conscience Self-Organizing Map to aid in the classification of particle signals from a dark matter experiment. We evaluate clusters extracted from the SOM for physics interpretation, label them and, by using the cluster labels, we demonstrate an improvement of accuracy over the currently used method. We also introduce a stopping criterion based on map quality to help shorten long SOM training.

1 Motivation

Over the past decade, the use of machine learning(ML) in particle physics has led to numerous new applications. Many of these applications depend on precise simulation-based forward modeling for supervised learning, which becomes feasible once experimental measurements mature [1]. However, since detectors in this field are typically custom-built prototypes, physicists need unsupervised learning techniques to help them understand their new data while commissioning the detectors. Our focus is on the detection of dark matter, a type of anomaly detection that employs detectors known for being notoriously difficult to model. To detect anomalies, robust signal classification in the detector is crucial for differentiating between artifacts and potential dark matter encounters.

Our work is inspired by experimental efforts that use instrumented vats of liquid xenon to terrestrially probe dark matter in underground laboratories ([2, 3]). Specifically, we discuss the case of the XENONnT experiment using publicly available information. In general, interactions in these detectors produce scintillation (S1) and ionization (S2) signals that are measured by photomultiplier tube (PMT) sensor arrays above and below the xenon vat. The stochastic nature of these signals, combined with a wide dynamic range, poses a significant challenge for signal classification suitable for rare event searches, especially when combined with unmodeled detector effects. Previous attempts at using ML for this classification have proven suboptimal due to these difficulties and the use of simulated data for training [4]. Here we use Conscience Self-Organizing Maps (CSOM) to improve signal classification over the currently used method.

*This work was supported in part by National Science Foundation Grants PHY-1719270, PHY-2112801, GFRP; Rice University IDEA Grant U50697, and Keck Foundation Gift 994453.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

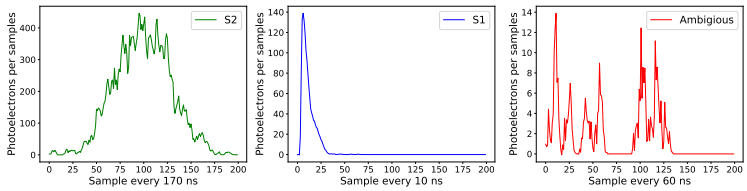

Fig. 1: An S2, S1, and an ambiguous signal. Left: An example of the typical waveform of an S2 signal with an almost Gaussian shape. Center: Canonical S1 signal, rises very quickly at the beginning and also drops off quickly. Right: Ambiguous signal.

2 The data

Generally, detectors like XENONnT record spatio-temporal data with time series data reported at 100 MHz from each sensor in the top and bottom PMT arrays. For classification purposes, we primarily analyze the waveforms, the summed time series data over all PMTs (Fig. 1), as the difference between S1 and S2 signal types (scintillation and ionization) is the time constant related to its production process. A 12-D SOM input vector is constructed from 10 waveform shape parameters, the area under the waveform (area) and the single variable 'area fraction top' (AFT) that encodes spatial information. The shape parameters correspond to the time it takes to reach 10%, 20%, ..., 100% of the total area of the waveform (the area deciles). The 'area fraction top' (AFT) represents the fraction of the area in the top array where signal is detected. In experiments, this set of inputs yielded the best classification results. We also use $\log_{10}$ scale for the area and deciles since this data spans seven orders of magnitude.

Training sets can generally consist of either simulation data (labeled with ground truth) or detector data of a radioactive isotope (e.g., ^{37}Ar), here we used 150,632 signals for training. To test classification results we simulated 20,000 signals labeled with ground truth using WFSIM [4], which represent 10,000 ^{37}Ar S1 signals and 10,000 small S2 (single electron (SE)) signals. These S1 and S2 subpopulations are the hardest to discriminate among all S1 and S2 subspecies, which presents a significantly increased classification challenge compared to previous works.

3 SOM-aided classification for XENONnT

To separate S1 from S2 signals [4] the algorithm currently used for XENONnT data (Straxen classification) applies several decision boundaries defined in a few 2-D feature subspaces (Fig. 2). The classification primarily relies on the rise-time (time between 10% and 50% of the waveform area) and area information. However, working with 2-D subspaces of n-D data may yield suboptimal results. With SOM learning we can separate the data clusters in the original n-D space and label the resulting clusters interactively as S1 or S2 signals using domain knowledge.

We apply a 40 x 40 CSOM [5] trained for 40 million steps, which is an educated guess for a sufficient number of learning steps based on previous ex-

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

periments. We use a CSOM instead of a Kohonen SOM [6] because the CSOM convergences faster and ensures a more faithful density matching [7]. We extract cluster boundaries using the CONN representation [8] at varying thresholds, validate and, if necessary, adjust the boundaries based on relevant physics parameters (such as the rise-time vs area space). For classification purposes, we group the 20 clusters found (S1 and S2 subspecies) into an S1 and an S2 supercluster by interactively labeling the SOM prototypes based on their features. This establishes a 2-class classification boundary.

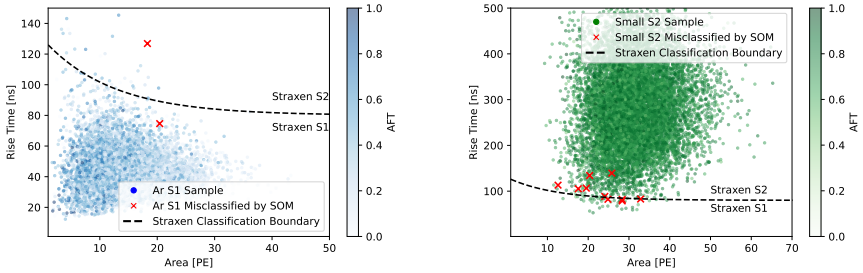

Fig. 2: Rise-time vs Area plots for the simulated ^{37}Ar S1-s (left) and small S2-s (right). The dashed curve indicates the Straxen classification boundary in this 2-D subspace. Samples that cross this boundary are misclassified by the current Straxen classification algorithm. Red 'x' symbols mark the data samples misclassified by the SOM-based approach, all other samples were predicted correctly.

We predict labels for simulated data by doing a full recall with the trained CSOM and assigning to each sample the label of the prototype to which it gets mapped. We use the aforementioned ^{37}Ar S1 and small S2 simulation data to evaluate our classification accuracy between these two subspecies.

Classification accuracy for small S2-s improved from 99.51% to 99.91%, and for ^{37}Ar S1-s it improved from 99.33% to 99.96% compared to the manual boundary currently used for classification. This increase in accuracy is significant because of the relatively high rate of small S2-s in the detector and their propensity for being misinterpreted as S1-s This is especially important for a rare event search where one of our main priorities is to avoid misclassifications between S1 and S2 signals that could lead to false dark matter detection.

The ability to work in n-D space allowed us to improve on a very stringent classification. Another benefit of this is the discovery of new S1 and S2 subspecies (not discussed here), which could be significant for our physics analysis.

4 A novel SOM stopping criterion

SOMs are powerful in expressing the structure of high-D, complex data, which facilitates precise detection of clusters and, in turn, accurate classification of new samples as in Section 3. However, SOM training can be long with no established stopping criteria vs the mapping quality for cluster extraction purposes. Measures of fit and topology preservation provide guidance for correct mapping ([7] and references therein, [9]) but nuanced measures are expensive and interpreta-

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

tion of less-than-perfect mapping —the degree of topology violation acceptable for accurate cluster detection— is difficult [7]. In this work we propose the Sample Migration (SaMi) Criterion, useful and cost-effective for this purpose.

SaMi expresses how settled the learning is in terms of the sample movement across prototypes between two learning stages. Here we assume the map is free of persistent folding on large scales. Denoting the receptive field of SOM prototype vector w_i, $i = 1, \cdots, M$ by $RF_i(ls1)$ and $RF_i(ls2)$ at learning steps $ls1$ and $ls2$, respectively, we compute the migration counts $gain_i(ls2)$ and $loss_i(ls2)$, the number of samples gained (lost) in $RF_i(ls2)$ compared to $RF_i(ls1)$. We call the points at which migration counts are computed SaMi points, and $mlag = ls2 - ls1$ the *migration lag* at $ls2$. $mlag$ can vary with $ls2$ ($mlag = mlag(ls2)$) as in this study, where $mlag(ls2)$ is equal to the variable stride between the SaMi point at $ls2$ and the previous SaMi point. Averaging over all M prototypes gives the mean migration counts $\overline{gain}(ls2)$ and $\overline{loss}(ls2)$ for gain and loss.

As samples become organized both the amount and the SOM lattice distance of moves should be diminishing. SaMi is intended to capture this. We discount moves within the neighborhood radius (r_{nb}) of *local violations*, l_{min}, as defined in [8]. $r_{nb} = l_{min}$ is the smallest radius in which all data-space (Voronoi) neighbor prototypes of w_i can be packed around w_i in the lattice. Within $r_{nb} = l_{min}$ topology violations are considered local, so sample movement can be excluded from the migration counts. In addition, we compute the migration statistics for $r_{nb} > 1$, i.e., excluding movements among immediate lattice neighbor prototypes.

These migration statistics are shown in the top row of Fig. 3 for three data sets of increasing size and complexity: the well-known Iris data with three clusters, a 6-D synthetic spectral image cube with eight distinct classes [7], and the XENON data (Section 2), containing 150 4-D, 16,384 6-D and 150,632 12-D samples, respectively. They were learned with 6 x 6, 15 x 15, and 40 x 40 CSOMs, and the l_{min} values are 1, 2, and 3, respectively. Points of learning rate drops are indicated by dashed vertical lines. (We omit elaboration on learning rate schedules for space considerations.)

Mean migration counts (Fig. 3, top row) sharply and steadily decrease after a learn step approx. 20 x the size of the respective data set (3K for Iris; 350K for the 6-D data, and 3M for XENON). Strikingly, after this point the vast majority of the sample migration occurs within the radius of local violations ($r_{nb} \leq l_{min}$): the movement to/from $r_{nb} > l_{min}$ (dot-dashed line) is only a small fraction of the total migration (upper dashed line). Moreover, the migration to/from $r_{nb} > 1$ is very close to the migration to/from $r_{nb} > l_{min}$, indicating that even within $r_{nb} \leq l_{min}$ most movements are across immediate neighbors. This is a significant phase change in map quality.

However, to recommend a stopping point we have to account for differences in the sizes of the data sets and migration lags, across and within data sets. For this we compute a normalized mean migration count, $n\_\overline{mcount}(ls2) = \overline{mcount}(ls2)/(mlag(ls2)/N)$, where $mcount$ is $gain$ or $loss$. $mlag(ls2)/N$ gives the number of times the entire data set is processed over the migration lag $mlag(ls2)$. This ignores other possible effects (such as stride size independent of

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

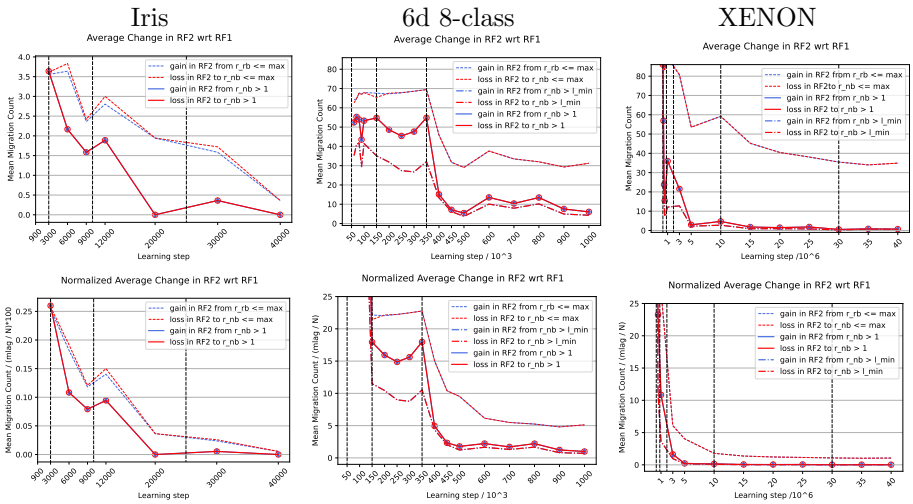

Fig. 3: Sample Migration (SaMi) statistics for three data sets. The upper (dashed) curves in each plot window are computed for the entire SOM lattice, $r_{nb} = max$ where max is the longest possible distance of two prototypes, using maximum distance. The solid and dot-dashed lines indicate the migration statistics for $r_{nb} > 1$ and $r_{nb} > l_{min}$, respectively. For the Iris data $l_{min} = 1$ thus the $r_{nb} > 1$ and $r_{nb} > l_{min}$ neighborhoods and curves coincide. **Top row:** Mean migration counts. **Bottom row:** Normalized migration counts, which compensate for the effects of sample size N, and migration lag $mlag$, within and across data sets.

$mlag(ls2)$, learning rates) but may be reasonable for a first exploratory study.

The normalized migration counts in the bottom row of Fig. 3 show emerging common trends among the three data sets. First, the point of sharp drop, and the alignment of the solid and dot-dashed curves are preserved from the non-normalized curves. Second, beyond the point of sharp drop all $r_{nb} > l_{min}$ curves fall within a range of 1-2% of the size of the respective data set computed as $prc\_n\_\overline{mcount}(ls2) = 100 * n\_\overline{mcount}(ls2)^*M \ / \ N$. This translates to 1.8% at $ls2 = 9K$ and below 1% at 20K for the Iris data; 2% at $ls2 \geq$500K for the 6-D data; 1.8 % at $ls2$=3M and 0.5% at 5M for XENON. This observation inspires a stopping recommendation at (or near) the SaMi point where $prc\_n\_\overline{mcount}(ls2) < 2\%$. While this cut-off is somewhat arbitrary, it provides a principled working hypothesis. We note that many sample moves in $r_{nb} > l_{min}$ could also be ignored because prototypes representing a cluster are in a close neighborhood at a mature learning stage, thus moves among them (often larger than l_{min}) are inconsequential for cluster capture. This makes the 2% threshold for our stopping criterion a loose upper bound.

We validate the proposed SaMi cut-off by comparing a benchmark cluster map with cluster maps extracted at each SaMi point between the cut-off and the end point, using the same prototype labeling as the benchmark. This is a stringent comparison (because an SOM could yield a perfect clustering even if the prototype labeling is somewhat different from the benchmark as prototypes may have moved) but we use it for this exploratory study to avoid interactive clustering of many SOMs. As a result, this validation may confirm safe stopping at a larger-than-necessary learning step. Uncertainties caused by the coarse

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

normalization and large stride between SaMi points prevent precise determination of a single learning point for stopping. However, the validation suggests that checking just a few SaMi points beyond the recommended cut-off may suffice for finding the point where the clustering quality matches the benchmark. For the three data sets here, validation confirms the stopping recommendations [9000,20000], [500,500], and [5M,15M], respectively. The width of these intervals —between 0% and 25% of the original learning steps— may be explained, besides differences in stride between SaMi points, by the crisp clusters in the 6-D synthetic data [7] vs intricate structure in the XENON data. While validation cautions against the possibly overoptimistic lower points in the above brackets, even with conservative use of the farther end of the brackets we save about 50% of the original learning steps in each case. Importantly, we see a promising trend worth deeper examination in more systematic and detailed experiments.

5 Conclusions and future work

We were able to train a CSOM on spatio-temporal physics data and show accuracy improvement of stringent classification over the current system on simulated data. This demonstrates that CSOMs could be a powerful tool for characterizing particle physics experiment data. We proposed the SaMi stopping criterion for SOM learning to shorten long production runs. While it provides a conservative estimate it can save 50% or more of the learning without compromising cluster accuracy. In future work we will strive to tighten this criterion based on insights from this exploratory study, including consideration of stride and learning rate effects, and weighting of sample moves by prototype connectivity [8].

References

[1] A. Radovic, M. Williams, D. Rousseau, et al. Machine Learning at the Energy and Intensity Frontiers of Particle Physics. *Nature*, 560(7716):41–48, 2018.

[2] E. Aprile, J. Aalbers, F. Agostini, et al. XENON1T Dark Matter Data Analysis: Signal and Background Models and Statistical Inference. *Phys. Rev. D*, 99:112009, Jun 2019.

[3] D.S. Akerib, C.W. Akerlof, D.Yu. Akimov, et al. The LUX-ZEPLIN (LZ) experiment. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 953:163047, 2020.

[4] XENON Collaboration, E. Aprile, K. Abe, S. Ahmed., et al. Detector Signal Characterization with a Bayesian Network in XENONnT. 4 2023.

[5] D. DeSieno. Adding a conscience to competitive learning. In *Proc. IEEE Int'l Conference on Neural Networks (ICNN), July 1988*, volume I, pages I–117–124, New York, 1988.

[6] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, Berlin Heidelberg New York, 1997.

[7] E. Merényi, K. Taşdemir, and L. Zhang. Learning highly structured manifolds: Harnessing the power of SOMs. In *Similarity-Based Clustering*, volume 5400 of *Lecture Notes in Computer Science*, pages 138–168. Springer Verlag, 2009.

[8] K. Taşdemir and E. Merényi. Exploiting data topology in visualization and clustering of Self-Organizing Maps. *IEEE Trans. on Neural Networks*, 20(4):549–562, 2009.

[9] Lutz Hamel. SOM quality measures: An efficient statistical approach. In *Proc. WSOM 2016, Advances in Self-Organizing Maps and Learning Vector Quantization, Advances in Intelligent Systems and Computing*, volume 428, pages 49–59. Springer, 2016.