ANALYZING BIOMEDICAL DATASETS WITH SYMBOLIC TREE ADAPTIVE RESONANCE THEORY: SUPPLEMENTARY MATERIAL

Sasha Petrenko
Department of Electrical and Computer Engineering
Missouri University of Science and Technology
Rolla, MO 65401
petrenkos@mst.edu

Daniel Hier
Department of Electrical and Computer Engineering
Missouri University of Science and Technology
Rolla, MO 65401
dhier@mst.edu

Mary Bone
University of Southeastern Norway (USN)
3616 Kongsberg, Norway
mary.bone@drmarybone.com

Tayo Obafemi-Ajayi
Engineering Program
Missouri State University
Springfield, MO 65897
tayoobafemiajayi@missouristate.edu

Erik J. Timspon
Honeywell Federal Manufacturing & Technologies
Kansas City, MO 64147
etimpson@kcnsc.doe.gov

William E. Marsh
Honeywell Federal Manufacturing & Technologies
Kansas City, MO 64147
wmarsh@kcnsc.doe.gov

Michael Speight
Honeywell Federal Manufacturing & Technologies
Kansas City, MO 64147

Donald C. Wunsch II
Department of Electrical and Computer Engineering
Missouri University of Science and Technology
Rolla, MO 65401
dwunsch@mst.edu

ABSTRACT

This article provides the written supplementary material to the paper titled *Analyzing Biomedical Datasets with Symbolic Tree Adaptive Resonance Theory*. This material concerns the evaluation of the START algorithm and its dual-vigilance and distributed dual-vigilance in comparison with existing algorithms, and it demonstrates the performance of all START variants on existing real-valued and categorical benchmark datasets. START is a multi-modal algorithm for both the clustering and the supervised learning of symbolic datasets containing statements under a grammar, so procedures for accommodating real-valued datasets are also described for benchmark evaluation.

Keywords adaptive resonance theory · biomedical data · categorical data · ontologies · knowledge graphs

1 Introduction

This article provides the supplementary material to the paper *Analyzing Biomedical Datasets with Symbolic Tree Adaptive Resonance Theory*. This material compares the START algorithm and its variants to existing algorithms in Section 2, and it demonstrates the performance of all START variants when tested against existing machine learning benchmark datasets in Section 3.

NSC-614-5761 dated 11/2023
Unclassified Unlimited Release

2 Comparison with Existing Methods

START is most directly comparable with Gram-ART for two important reasons: Gram-ART is the first and, prior to START, only ART-based symbolic data clustering algorithm, and the design of START uses Gram-ART as a basis with some important modifications. START differs from Gram-ART in the following ways:

1. The original formulation of Gram-ART uses an ART match function as both a match and activation function for evaluation of prototype trees upon the samples, whereas START utilizes separate ART activation and match functions in the vein of FuzzyART to distinguish the WTA competition dynamics. The activation function is evaluated for all prototypes, which then undergo a WTA competition for satisfying the match criteria.
2. START prototype trees do not contain terminal symbol positions, instead encoding encountered symbols entirely in the PMFs of the non-terminal positions corresponding to the production rules of those terminals.
3. The original meta-learning of genetic algorithms application of Gram-ART benefited from penalizing structural dissimilarity between statements, so it did not have a mechanism for the structural evolution of prototypes. On the other hand, START adds the ability to grow prototype structures through the design decision of encoding purely non-terminal positions. If a statement satisfies the vigilance criteria of a given existing prototype, new nodes denoting non-terminal symbols are inserted into the prototype tree where necessary to encode the occurrence of terminal symbols at and below novel non-terminal positions along the prototype tree.
4. Dual-Vigilance (DVSTART) and Distributed Dual-Vigilance (DDVSTART) variants are additionally derived for the symbolic tree prototype representations of START following recent advancements in the ART literature [1, 2].
5. A simplified supervised modification is defined for all START variants in the manner of Simplified FuzzyARTMAP [3, 4]; a cluster-label mapping and forced mismatch procedure enables the supervised training and performance testing of START variants when integered or ordinal-encoded labels are available, such as in the benchmark machine learning datasets used in Section 3.

3 Benchmark Evaluation

START is a multi-modal algorithm for the clustering and supervised training of symbolic datasets in the form of statements under a grammar. It is developed from It has an original START variant, dual-vigilance DV-START variant, and distributed dual-vigilance DDV-START variant. It is formulated to work with symbolic data, but it can be used for both the clustering and the supervised training and testing of real-valued datasets when a binning procedure is applied, which is outlined in Section 3.1.

Gram-ART is originally verified upon a discretized version of the UCI Iris dataset, the UCI Mushroom dataset, and the UCI Unix User dataset [5, 6, 7, 8]. For comparison, START is evaluated upon the following open-source machine learning benchmark datasets with existing labels: a set of real-valued clustering benchmark datasets [9, 10], the categorical UCI mushroom dataset [7], and a categorical lung cancer patient dataset [11].

3.1 Method

Because the benchmark datasets such as the Iris flower dataset elements are real-valued, each feature is range-normalized and binned into a set of terminal symbols representing each bin. The resulting grammar from this symbolic binning procedure is typified by the discretized Iris grammar in Table S1.

3.2 Results

Table S2 demonstrates the results of training and testing the supervised mode of each START variant. Real-valued datasets utilize the binning procedure outlined in Section 3.1 with number of bins parameter $M = 10$. Hyperparameters of this discretization process are not varied in the following results to minimize the total number of hyperparameters to vary in the study. START has one hyperparameter with the vigilance parameter ρ, while DV-START and DDV-START have both lower-bound and upper-bound vigilance parameters ρ_{lb} and ρ_{ub}, respectively. DDV-START also has the additional HAC linkage method as a hyperparameter.

An initial hyperparameter sweep is done for each combination of START variant hyperparameter, dataset, and varying training dataset shuffle random seeds; the same 80% training and 20% testing split is maintained or each dataset, with only the input presentation order of the training samples being varied. Each variant is trained in supervised mode with

Listing 1: Discretized Iris dataset grammar illustrating the symbolic binning procedure of real-valued data used to evaluate START and Gram-ART [5]. Each statement S consists of four terminal symbols sampled from each original feature dimension. Each feature dimension is cast into a non-terminal symbol in set N that is realized by a set of terminal symbols T representing a range of real-valued bins, where the number of bins M is a user-selected hyperparameter. For brevity, ellipses in each terminal symbol field represent the set of bins between the first and last elements of the binning procedure. Iris feature dimensions are abbreviated with their initialisms (e.g., "SL" = "Sepal Length"). Production rules P follow the extended Bachus-Naur form (EBNF) notation mapping non-terminal symbols to the set of terminals belonging to the feature dimension bins.

$$\langle S \rangle ::= \langle SL \rangle \langle SW \rangle \langle PL \rangle \langle PW \rangle;$$

$$\langle SL \rangle ::= \texttt{'SL1'} \mid \ldots \mid \texttt{'SLM'};$$

$$\langle SW \rangle ::= \texttt{'SW1'} \mid \ldots \mid \texttt{'SWM'};$$

$$\langle PL \rangle ::= \texttt{'PL1'} \mid \ldots \mid \texttt{'PLM'};$$

$$\langle PW \rangle ::= \texttt{'PW1'} \mid \ldots \mid \texttt{'PWM'};$$

Table S1: Hyperparameters for each START variant during supervised train/test evaluation. Each START variant varies in its vigilance parameter(s), with DDV-START additionally varying its HAC linkage method as described in the original paper. Each variant additionally varies the presentation order of the training samples by varying the random seed value prior to shuffling training samples before training. The set of training and testing samples are maintained from an initial $80\%/20\%$ split with only the training sample presentation order being varied to calculate testing performance statistics.

Variant	Hyperparameters
START	ρ, `seed`
DV-START	ρ_{lb}, ρ_{ub}, `seed`
DDV-START	ρ_{lb}, ρ_{ub}, `linkage`, `seed`

the labels provided in each dataset, and a single training pass is done through the training samples (i.e., each sample is presented only once sequentially with no retraining).

Performances of each hyperparameter in the initial sweep are averaged across training presentation orders, and best-performing hyperparameters are selected for each variant-dataset combination for a final hyperparameter sweep (Table S2). Training presentation order is varied with 1000 different random shuffle seeds to generate statistics of performance as a function of training data presentation order.

All varied hyperparameters are listed for each variant in Table S1 for both the initial hyperparameter selection sweep and for the final results statistics seen in Table S2. Performances in the final results sweep are represented as means and variances of top-1 testing accuracy across all 1000 training/testing iterations for each variant-dataset combination.

All three variants show similar results in performance on both the real-valued and categorical datasets used for evaluation in both degree of accuracy and in the statistics of their accuracies with minor variation, achieving near-perfect testing accuracies in several instances. Variation in top-performing variant for each dataset is to a small enough degree that can be attributed to dependence on hyperparameter selection from the initial sweep. High performance on the real-valued datasets is generally unexpected due to the discretization process as outlined in Section 3.1 that is necessary to adapt the datasets to the symbolic formulation of START and its variants, but consistently competitive performances are found nonetheless with minimal variation. START inherits the relative resilience to input presentation order dependence of ART-based algorithms while at the same time requiring only a training single pass to achieve these performances. This ART-based formulation of START and its variants also inherits the ability to continue learning new categories in a lifelong and continual learning manner that results in high testing performances after only a single training pass.

START and its variants also test with high accuracies on the evaluated symbolic datasets, verifying their ability to accurately learn directly on purely categorical datasets without specialized encoding schemes. These performance values are indeed expected because the categorical benchmark datasets used are interpreted as symbolic datasets with fixed-length statements, which underutilizes START's capacity to learn arbitrarily long symbolic statements when interpreted as parse trees under a formal grammar, such as the protein-disease dataset studied in the original paper.

Table S2: Performance statistics of the supervised implementations of each START variant derived in the original paper on a set of benchmark real-valued and categorical machine learning datasets. The mean and variance $\mu \pm \sigma^2$ of the testing performance is presented to 4 decimal places across 1000 training sample presentation orders for the START algorithm and its Dual-Vigilance (DV-START) and Distributed Dual-Vigilance (DDV-START) variants [1, 2, 4]. Best-performing hyperparameters are selected from an initial sweep of variant-dataset combinations. A fixed set of an $80\%/20\%$ training/testing split is maintained for all iterations, and the training set is randomly shuffled 1000 times to generate testing performance statistics for each variant-dataset combination. Supervised mode is used for each START variant, and the training data is presented in a single pass with no retraining. The best average performance for each dataset is bolded. Just as in [2], a set of real-valued datasets from [9, 10] are used with the discretization procedure outlined in Section 3.1, while the categorical UCI mushroom dataset [7] and categorical lung cancer patient dataset [11] are trained upon directly in original symbolic form. Real-valued benchmark machine learning datasets are listed first, and the `LungCancer` and UCI `Mushroom` categorical datasets are listed thereafter.

Dataset	START	DV-START	DDV-START
Aggregation	0.9860 ± 0.0001	$\mathbf{0.9861} \pm 0.0001$	0.9860 ± 0.0001
Compound	0.9035 ± 0.0009	0.9051 ± 0.0009	$\mathbf{0.9056} \pm 0.0009$
R15	0.8982 ± 0.0017	0.8983 ± 0.0016	$\mathbf{0.8984} \pm 0.0017$
Flame	$\mathbf{0.9474} \pm 0.0011$	0.9482 ± 0.0010	0.9473 ± 0.0010
Jain	$\mathbf{0.9941} \pm 0.0001$	0.9938 ± 0.0001	0.9939 ± 0.0001
PathBased	$\mathbf{0.9520} \pm 0.0007$	0.9507 ± 0.0007	0.9505 ± 0.0008
Spiral	$\mathbf{0.9643} \pm 0.0006$	0.9635 ± 0.0007	0.9629 ± 0.0007
Face	0.9732 ± 0.0004	$\mathbf{0.9743} \pm 0.0003$	0.9734 ± 0.0004
Flag	$\mathbf{0.9998} \pm 0.0000$	0.9997 ± 0.0000	0.9996 ± 0.0000
Halfring	$\mathbf{0.9927} \pm 0.0001$	0.9922 ± 0.0002	$\mathbf{0.9927} \pm 0.0001$
Iris	0.8871 ± 0.0031	$\mathbf{0.9094} \pm 0.0020$	0.8835 ± 0.0032
Moon	$\mathbf{0.9886} \pm 0.0001$	0.9877 ± 0.0001	0.9885 ± 0.0001
Ring	0.9999 ± 0.0000	$\mathbf{1.0000} \pm 0.0000$	$\mathbf{1.0000} \pm 0.0000$
Spiral	0.9309 ± 0.0020	0.9314 ± 0.0021	$\mathbf{0.9344} \pm 0.0020$
Wave	$\mathbf{0.9974} \pm 0.0001$	0.9967 ± 0.0001	0.9967 ± 0.0001
Wine	0.8840 ± 0.0021	$\mathbf{0.8962} \pm 0.0016$	0.8771 ± 0.0023
LungCancer	0.9998 ± 0.0000	$\mathbf{0.9999} \pm 0.0000$	0.9998 ± 0.0000
Mushroom	0.9146 ± 0.0006	$\mathbf{0.9159} \pm 0.0006$	0.9151 ± 0.0007

Testing performance on the `LungCancer` and `UCI Mushroom` datasets demonstrates at a minimum the capacity of START and its variants to learn rudimentary symbolic datasets in supervised mode.

Funding

This work is funded by the Department of Energy's Kansas City National Security Campus, operated by Honeywell Federal Manufacturing & Technologies, LLC, under contract number DE-NA0002839.

References

[1] L. E. Brito da Silva, I. Elnabarawy, D. C. Wunsch, L. Enzo, I. Elnabarawy, D. C. Wunsch, L. E. Brito da Silva, I. Elnabarawy, D. C. Wunsch, L. Enzo, I. Elnabarawy, D. C. Wunsch, L. E. Brito da Silva, I. Elnabarawy, and D. C. Wunsch, "Dual vigilance fuzzy adaptive resonance theory," *Neural Networks*, vol. 109, pp. 1–5, 2019.

[2] L. E. Brito da Silva, I. Elnabarawy, and D. C. Wunsch, "Distributed dual vigilance fuzzy adaptive resonance theory learns online, retrieves arbitrarily-shaped clusters, and mitigates order dependence," *Neural Networks*, vol. 121, pp. 208–228, 2020.

[3] T. Kasuba, "Simplified fuzzy artmap, ai expert," 1993.

[4] L. E. Brito da Silva, I. Elnabarawy, and D. C. Wunsch, "A Survey of Adaptive Resonance Theory Neural Network Models for Engineering Applications," *Neural Networks*, vol. 120, no. xxxx, pp. 167–203, 2019.

[5] R. J. Meuth, *Adaptive multi-vehicle mission planning for search area coverage*. PhD thesis, Missouri University of Science and Technology, 2007.

[6] R. A. Fisher, "Iris." UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C56C76.

[7] "Mushroom." UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C5959T.

[8] T. Lane, "UNIX User Data." UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5302K.

[9] N. Ilc, "Datasets package," 06 2013.

[10] P. Fränti and S. Sieranoja, "K-means properties on six clustering benchmark datasets," 2018.

[11] A. S. Ahmad and A. M. Mayya, "A new tool to predict lung cancer based on risk factors," *Heliyon*, vol. 6, p. e03402, feb 2020.