Spring 2015

Fuzzy adaptive resonance theory: Applications and extensions

Clayton Parker Smith

FUZZY ADAPTIVE RESONANCE THEORY:

APPLICATIONS AND EXTENTIONS

by

Clayton Parker Smith

A THESIS

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE IN COMPUTER ENGINEERING

2015

Approved by

Donald C. Wunsch II, Advisor
R. Joe Stanley
Gayla R. Olbricht

PUBLICATION THESIS OPTION

This thesis has been prepared in the form of two conference papers formatted according to the university specifications. The first paper consisting of pages 11-20 will be submitted to the International Joint Conference on Neural Networks. The second paper consisting of pages 21-30 will also be submitted to the International Joint Conference on Neural Networks.

ABSTRACT

Adaptive Resonance Theory, ART, is a powerful clustering tool for learning arbitrary patterns in a self-organizing manner. In this research, two papers are presented that examine the extensibility and applications of ART. The first paper examines a means to boost ART performance by assigning each cluster a vigilance value, instead of a single value for the whole ART module. A Particle Swarm Optimization technique is used to search for desirable vigilance values. In the second paper, it is shown how ART, and clustering in general, can be a useful tool in preprocessing time series data. Clustering quantization attempts to meaningfully group data for preprocessing purposes, and improves results over the absence of quantization with statistical significance.

ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Donald Wunsch, for his support and guidance. He has opened many doors for me since he accepted me as an undergraduate advisee and, later, research assistant. Through his patience, I development my professional and technical skills. I would like to acknowledge the contributions from my M.S. committee, Professors R. Joe Stanley and Gayla Olbricht, who helped strengthen this thesis.

This work would not have been possible without the financial support from the M.K. Finley Endowment and the Chancellor's Fellowship.

I would also like to thank Sherri Smith and Brynne Coleman for supporting my journey through college these past seven years.

TABLE OF CONTENTS

Page

LIST OF ILLUSTRATIONS

Page

LIST OF TABLES

Page

1. INTRODUCTION

1.1. OVERVIEW

Today's need for data analytic techniques is great. Biology has been the muse for data processing and optimization. Numerous methods created during the latter half of the 20th century were biologically inspired, (e.g., artificial neural networks, particle swarms, fuzzy logic, genetic and evolutionary computing, and artificial immune systems).

Biologically-inspired machine learning methods have seen success in linear and nonlinear function approximations, data processing, and classification. Applications include filtering, adaptive control, pattern recognition, and pattern discovery. The utility in these applications were evident across many disciplines.

Machine learning has been deployed across many disciplines, (e.g., psychology, neuroscience, statistics, etc). Cognitive psychology has devoted itself to theories of learning. Socrates was one of the first to study the learning process, noting that knowledge comes from within [13-14]. Pavlov demonstrated that dogs could be conditioned to salivate via a reinforcement signal from a bell [18]. Several studies have been conducted to understand the brain's primitive functions, its ability to group objects and concepts, and its ability to think abstractly [15-17].

Clustering is one of these primitive functions the brain performs. Gail Carpenter and Stephen Grossberg developed theories on not only clustering, but also how the brain learns [1-4]. They created Adaptive Resonance Theory (ART). This concept utilizes resonance as part of a learning theory.

Adaptive Resonance Theory has been used successfully as a powerful data clustering tool. It can learn arbitrary patterns quickly in a self organizing way. To briefly

compare and contrast with k-Means clustering [5-9], ART is a parameterized algorithm. In k-means, the number of clusters must be specified a priori, while ART has a vigilance threshold. This threshold allows for the creation of new clusters in real-time. The vigilance threshold also determines how tight or loose the recovered clusters are.

1.2. LEARNING PARADIGMS

Most machine learning methodologies, particularly in neural networks, can be classified into one of three main learning paradigms. They are: supervised learning, unsupervised learning, and reinforcement learning. Several other paradigms exist, but they are, primarily, based on one of these three (e.g. semi-supervised learning, which hybridizes the ideas of supervised and unsupervised learning).

1.2.1. Supervised Learning. Supervised learning is synonymous with having a teaching or training signal, or oracle, that has a perfect knowledge of the defined task. It knows the answer to arbitrary inputs into the system and can evaluate the response with a desired response. A machine learning system utilizing this learning paradigm would be able to correct itself by taking into account the disparity between its response and the desired response. The system would be guiding itself towards a minima of error. Teaching a system to learn the response behavior of a quadratic would illustrate this paradigm.

1.2.2. Unsupervised Learning. Unsupervised learning is similar to allowing the machine learning algorithm to take care of itself. The learning paradigm relies heavily on both the mathematical and statistical properties associated with the problem domain. These properties are used, ideally, to glean meaningful knowledge from the relational

aspect of the problem applied. This concept can be illustrated by grouping blocks by shape or size, each being a measure of similarity.

1.2.3. Reinforcement Learning. A number of problems with complex dynamics make supervised learning useless. In these situations, the computational burden of calculating the appropriate response for any arbitrary input becomes too great. Reinforcement learning is ideal in these instances. This approach is well-suited when explicit output recommendations are not available or are only available a minority of the time. Particularly when there are no explicit recommendations, an excellent substitute for such recommendations is a cost function. Reinforcement learning can be thought of as the process of causing a cost function to replace error signals that would have come from a teacher if one were available. A control problem (e.g., a cart balancing a pole on a 2-D track) is one example of a good use of reinforcement learning.

1.3. CLUSTERING[1]

Clustering is a powerful methodology for data analysis that humans perform on a daily basis. People are constantly bombarded with information as they move about their day. This information becomes processed, organized, and examined. Descriptive features can be identified when a new object or phenomenon is encountered. When comparing these features to known objects or phenomena, the unknown can become known. Humans have an unquantifiably large corpus of data to work with. This information is used to gain knowledge and understanding about the world around them.

[1] Section 1.3 is derived from [10]

The ability to group, or classify, data and examine emergent patterns is at the forefront of data acquisition. Data, when grouped together, is expected to exhibit similar properties under certain criteria. For a system to learn the emergent characteristics in data, it must either create labels autonomously or adjust system parameters to recognize known labels implicitly.

Class labels are known in supervised classifications. From a set data vectors, denoted as $x \in \Re^d$, where d is the dimensionality of the input space, a mapping exists to a finite set of discrete class labels, designated as $y \in 1,...,C$, where C is the total number of classes [10]. The system can then be modeled as

$$y_i = f(x_i, w) \tag{1.1}$$

where w is defined as the vector of the system parameters and i denotes an arbitrary input. The system parameters can be iteratively updated to minimize the overall system error on a finite sample of output mapped data vectors, $i=1,...,n$, where n is the total number of samples. The system can perform functionally as a classifier when the system either converges to an acceptable level of system error or reaches a prescribed number of update iterations.

Data labels are unknown in unsupervised classification. Unsupervised classification has been referred to as clustering or exploratory data analysis. Clustering methods attempt to discover some hidden, underlying structure from within a finite set of data vectors, denoted as $x \in \Re^d$, where d is the dimensionality of the input space. Most clustering algorithms fall into one of two categories: hierarchical and partitional.

Hierarchical clustering is split into two branches: agglomerative and divisive, see Figure 1.1. Agglomerative clustering builds groups from the bottom-up, beginning at

individual data points. Divisive clustering takes a top-down approach and looks for logical splits.

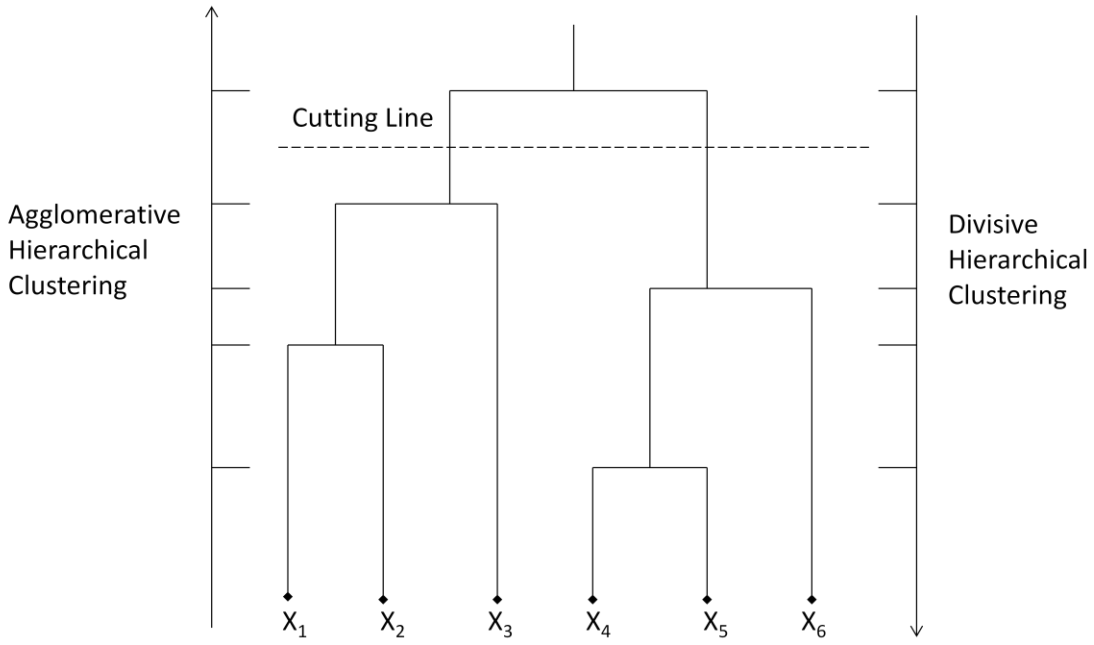

Figure 1.1. Agglomerative and Divisive Hierarchical Clustering

Partitional clustering can be either hard or fuzzy, see Figure 1.2. Hard partitions form crisp boundaries where data vectors definitively either belong or do not belong to a cluster. Fuzzy partitions form fuzzy boundaries where data vectors have a degree of membership to different clusters. This fuzzy membership is based on a fuzzy membership function. The fuzzy membership function's formulation can be based on a similarity measure though it is ultimately defined by the practitioner.

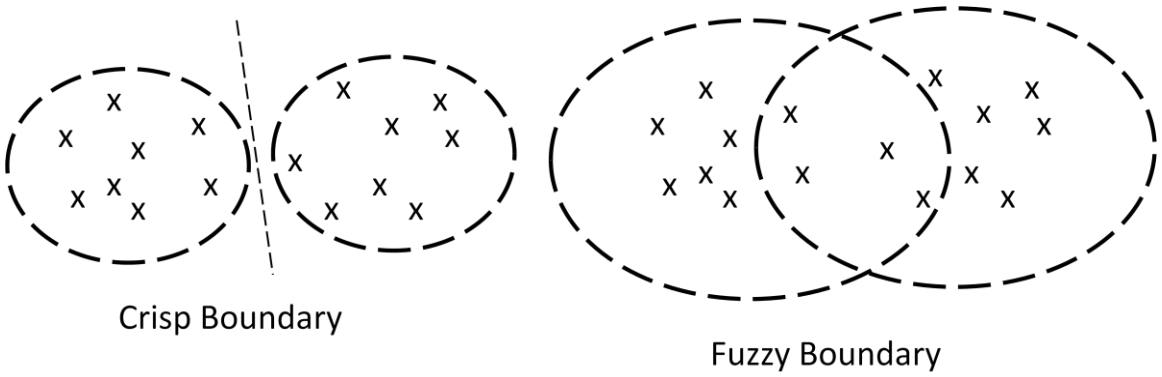

Figure 1.2. Hard and Fuzzy Partitional Clustering

1.4. VALIDATION MEASURES

Methods must be established to not only determine the quality of the clustering results, but also validate the clustering algorithm. Thus, cluster validation indexes have been researched a great deal [11-12,21-22]. All methods will fall into one of three categories; external criterion methods, internal criterion methods, and relative criterion methods. Several studies combined these three methods into two [11-12].

External criterion measures will generally compare clustering results, *C*, with some a priori knowledge. In some cases, this could be the ground truth; in others, it may be comparing it to another result. Internal criterion measures will generally include an examination of the clustering result's internal structure. Both the compactness of and the separation from the clusters with respect to one another would be investigated. The diversity of this evaluation method stems from the numerous ways in which compactness and separation can be quantified. Relative criterion measures will generally compare the clustering results *C* with other clustering results. This could take the form of comparing

the results using different cluster algorithm parameters and examining the change in the corresponding external or internal measures.

1.5. MANIPULATING ART

ART is based off of neural networks and, therefore, has a simple extensible architecture, see Figure 1.3. Its self-organizing property grants a degree of autonomy that is particularly useful when compared to methods without this property. ART is a cognitive theory for learning [1-4,9]. Its architecture is a framework for the learning theory. As a framework, pieces can be removed and new pieces added in. New systems can be built from the old [2,9,19-20].

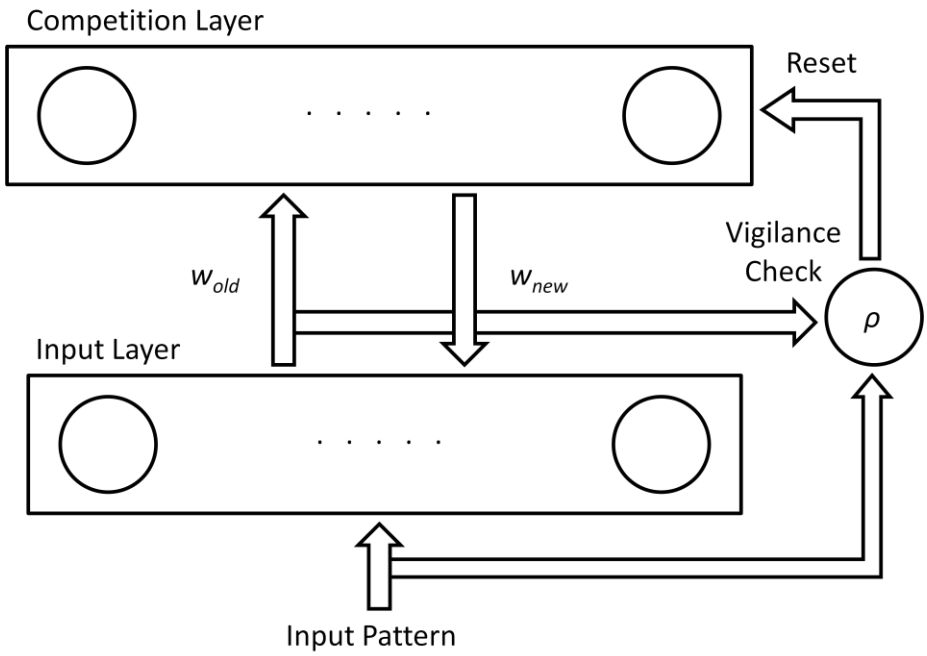

Figure 1.3. Adaptive Resonance Theory Framework

The original ART implementation could only handle binary input data [1]. While there are many problems that can be formulated in a discrete manner, much of the world regularly operates in analog. Fuzzy logic provided the extension necessary to expand ART into the continuous domain [2]. Between its discrete and analog forms, ART has a lot to offer in engineering applications [8].

ART functions primarily in an unsupervised manner. There are drawbacks with this autonomous learning. Natural partitions that are sparse may be needlessly broken up into multiple clusters. An extension to ART was developed to map these unnecessary divisions back to their natural partitions [9]. This changes the nature of ART from an unsupervised learning method, to a supervised learning method.

This is only a sample of the many extensions that have been developed for ART. The extensions presented are meant to show the utility and extensibility of ART. This provides a foundation for the rest of this thesis.

1.6. CLOSING NOTES

This research was focused on manipulating ART. The first paper in this work includes a discussion on the use of different vigilance values for each recovered cluster rather than a blanket vigilance threshold for the entire ART module. This is done by employing a particle swarm technique for the vigilance search. The second paper discussed the use of clustering techniques (e.g., ART) to preprocess and cluster sequential data for prediction purposes.

REFERENCES

[1] G Carpenter, S Grossberg. The ART of Adaptive Pattern Recognition by a Self-organizing Neural Network. IEEE Computer, vol 21, part 3, pp 77-88, 1988.

[2] G Carpenter, S Grossberg, D Rosen. Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System. Neural Networks, vol 4, pp759-771, 1991.

[3] S Grossberg. How does the brain build a cognitive code. Psychological Review, pp 1-51, 1980.

[4] S Grossberg. Competitive learning: From interactive activation to adaptive resonance. Cognitive Science, vol 11, pp 23-63, 1987.

[5] J MacQueen. Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkley Symposium on Mathematical Statistics and Probability, Statistics, vol 1, University of California Press, 1967.

[6] B Moore, M Fogaca, A Kramer. Characterizing the Error Function of a Neural Network. 2nd Symposium on the Frontiers of Massively Parallel Computation, pp 49-57, 1988.

[7] R Xu, D Wunsch. Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, vol 16, no 3, pp 645-678, 2005.

[8] D Wunsch. ART Properties of Interest in Engineering Applications. IJCNN, pp 3380-3383, 2009.

[9] G Carpenter. Default ARTMAP. Proceedings of International Joint Conference on Neural Networks, vol 2, pp 1396-1401, 2003.

[10] R Xu, D Wunsch. Clustering. Wiley, 2009.

[11] J Wu, H Xiong, J Chen, W Zhou. A Generalization of Proximity Functions for K-means. Proceedings of the 2007 IEEE International Conference on Data Mining, 2007.

[12] H Xiong, J Wu, J Chen. K-means Clustering versus Validation Measures: A data distribution perspective. Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining, pp 779-784, 2007.

[13] M Ference, Shirley Booth. Learning and Awareness. Lawrence Erlbaum Associates, 1997.

[14] B Hergenhahn. An Introduction to the History of Psychology. Belmont: Thompson Learning, 2005.

[15] P Gray. Psychology, 5th ed. New York: Worth, p 281, 2006.

[16] J Wolfe, K Kluender, D Levi, L Bartoshuk, R Herz, R Klatzky, S Lederman. Gestalt Grouping Principles. Sensation and Perception, 2nd ed. Sinauer Associates, 2008.

[17] J Banerjee. Gestalt Theory of Perception. Encyclopaedic Dictionary of Psychological Term, M.D. Publications Pvt. Ltd., pp 107-109, 1994.

[18] N Sheehy, A Chapman, W Conroy. Ivan Petrovich Pavlov. Biographical Dictionary of Psychology, Routledge, 2002.

[19] C-J Lin, C-T Lin. An ART-based Fuzzy Adaptive Learning Control Network. IEEE Transactions on Fuzzy Systems, vol 5, no 4, pp477-496, 1997.

[20] L Meng, A H Tan, D Wunsch. Vigilance Adaptation in Adaptive Resonance Theory. IEEE International Joint Conference on Neural Networks, pp 1-7, 2013.

[21] J Bezdek, N Pal. Some New Indexes of Cluster Validity. IEEE Transaction on Systems, Man, and Cybernetics, Part B: Cybernetics, vol 28, no 3, pp301-315, 1998.

[22] Y Liu, Z Li, H Xiong, X Gao, J Wu. Understanding of Internal Clustering Validation Measures. IEEE 10th International Conference on Data Mining, pp911-916, 2010.

PAPER

I. PARTICLE SWARM OPTIMIZATION IN AN ADAPTIVE RESONANCE FRAMEWORK

1.1. ABSTRACT

A Particle Swarm Optimization (PSO) technique, in conjunction with Fuzzy Adaptive Resonance Theory (ART), was implemented to adapt vigilance values to appropriately encompass the disparity in data sparsity. Gaining the ability to optimize a vigilance threshold over each cluster as it is created is useful because not all conceivable clusters have the same sparsity from the cluster centroid. Instead of selecting a single vigilance threshold, a metric for the PSO to optimize on must be selected. This trades one design decision for another. The performance gain, however, motivates the tradeoff in certain applications.

1.2. INTRODUCTION

Adaptive Resonance Theory (ART) has been used successfully in a variety of applications [17-20]. A number of other clustering methods require the user to specify the number of clusters desired a-priori. Adaptive Resonance Theory, however, only requires that the user set a vigilance threshold. This threshold determines how tight or loose clusters are, allowing ART to create new clusters autonomously.

One of the primary disadvantages of the vigilance threshold is that it applies to all possible clusters. Two clusters, in which one is tightly packed and the other is large and loose, can be easily imagined. A single vigilance value would not achieve high fidelity

for each cluster. This motivates the idea of using a different vigilance threshold for each cluster, e.g. [4]. The problem then becomes determining the vigilance for each cluster, as it is created.

As an alternative to [4], Particle Swarm Optimization (PSO), another biologically inspired machine learning method, is well-suited for this task. Several studies combined PSO with clustering methods (e.g., ART) [1,3]. Balancing the dichotomy of exploration and exploitation, PSO assists in searching for candidate vigilance thresholds.

This paper is organized into four sections. The methods employed, PSO, ART, and their combination, are discussed in Section 2. Section 3 is focused on the data used, the experiments conducted, and the results gathered. Section 4 concludes the paper.

1.3. THEORY

1.3.1. Fuzzy Adaptive Resonance Theory. Fuzzy Adaptive Resonance Theory. ART, is a learning theory. It overcomes the stability-plasticity dilemma and can learn arbitrary input patterns in a stable, fast, and self-organizing way [12,13,15,16]. A particularly useful variant of ART is Fuzzy ART [13]. The details reviewed below are useful for understanding how vigilance was modified in this study.

The architecture for Fuzzy ART has two layers: the F1 Layer and the F2 Layer. Normalized input patterns, comprising the F1 layer, are fed through a weight matrix, which acts as a category template. Category choices are calculated for each F2 category against the input vector:

$$T_j = \frac{\left| x \wedge w_j \right|}{\alpha + \left| w_j \right|} \tag{1.1}$$

where $\wedge$ is the fuzzy AND operator defined by

$$(x \wedge y)_i = \min(x_i, y_i) \tag{1.2}$$

and α is a choice parameter that is used to break ties.

In a winner-take-all fashion, the highest category choice is taken. The category match equation is used to compare the winning node to the vigilance threshold:

$$\rho \leq \frac{|x \wedge w_j|}{|x|} \tag{1.3}$$

If the node is classified as a match, that input pattern is mapped to the selected node. If the node is not a match, that node is turned off via a reset mechanism, and a new competition in the F2 layer takes place. The cluster mapping is built as each input pattern is matched to a node. The vigilance threshold greatly affects the ART network's performance, as it determines the criteria for the "goodness" of the match.

1.3.2. Particle Swarm Optimization. Particle Swarm Optimization is a technique by which a swarm of simple agents traverse an n-dimensional search space, attempting to find global minima/maxima. It attempts to balance the dichotomy of exploitation and exploration [5].

In PSO, a number of particles are initialized randomly within the search space with a random velocity. The particle's position at each iteration is evaluated according to a fitness function. Each particle's best position is noted, and the swarm's best position is determined. A new velocity is then calculated. This takes into account its previous velocity, weighted towards its best position and the global best position. The velocity update can be calculated as

$$v_{t+1} = \omega * v_t + \varphi_p * r_p * (p_t - x_t) + \varphi_g * r_g * (g_t - x_t) \tag{1.4}$$

where v is the particle's velocity, x is the particle's position, p is the particle's best position, g is the global best position, ω is a weighting term, φ is a weighting term with

respect to both the particle's best and global best, and *r* is a random applied weight that shifts the balance between the particle's best and the global best position.

1.3.3. ART - PSO Hybrid. The ART category creation event is the ideal place in the algorithm for a PSO hybridization. When ART creates a new category, the vigilance vector is then incremented and a new swarm is initialized to optimize ART's performance.

This extends ART with vigilance thresholds for each clusters, optimizing each threshold to its cluster. This hybridization attempts to make ART responsive to variations in cluster compactness. The datasets that include both tight and loose clusters should benefit from this approach.

1.3.4. Validation Indexes. Four validation indexes were chosen for the PSO to optimize: classification accuracy, the Rand index, the Silhouette index, and the Dunn index.

The easiest index to define is accuracy. Accuracy is simply the ratio of correctly classified data elements over the total number of data elements.

The Rand Index requires the computation of a confusion matrix. A true positive (TP) corresponds to two similar data points being assigned the same cluster. A true negative (TN) corresponds to two dissimilar data points being assigned to different clusters. A false positive (FP) corresponds to two dissimilar data points being assigned to the same cluster. A false negative (FN) corresponds to two similar data points being assigned to different clusters. With these four variables in mind, we can define the Rand Index by

$$R = \frac{TP + TN}{TP + FP + FN + TN} \tag{1.5}$$

The Silhouette index examines the relationship that exists between of the clustering results and the data that goes into it. It takes into account the cohesion within a cluster and the dissimilarity with other clusters.

Consider each datum *i* and, further, let *a(i)* be the average dissimilarity between *i* and all other data within the same cluster. This depiction gives insight to the cluster's cohesion. Let *b(i)* be the smallest average dissimilarity between *i* and every other cluster to which *i* is not a member. The silhouette index can then be defined as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{1.6}$$

The Dunn index examines both the compactness and the separation of the recovered clusters. Formulating distance measures, between clusters, when left up to the practitioner, can have a great impact on the results. The distance between clusters will be defined as the smallest distance between a pair of points that belong to each cluster. The diameter, or size, of a clusters is the largest distance between two of its members. The Dunn index is defined as

$$D(K) = \min_{i=1,\ldots,K} \left(\min_{j=i+1,\ldots,K} \left(\frac{dist(C_i, C_j)}{\max_{i=1,\ldots,K} diam(C_l)} \right) \right) \tag{1.7}$$

where K is the number of clusters.

1.4. DATA, EXPERIMENTS, AND RESULTS

Three datasets were chosen to test the efficacy of this layered adaptability approach to ART. The Iris, Wine, and Wisconsin Breast Cancer datasets within the UCI Repository [10] are common benchmark datasets that are often used to test clustering algorithms. The Iris dataset contains three classes, two of which are partially inter-

mixed, with four descriptors: petal length, petal width, stamen length, and stamen width. The Wine dataset contains three classes and thirteen descriptors. The Wisconsin Breast Cancer (WBC) dataset contains two classes and nine descriptors.

Four metrics were chosen for the PSO to optimize, two external and two internal measures. The Accuracy and Rand indices were chosen because they utilize the ground truth of the dataset in question in their calculation. The Silhouette and Dunn indices were chosen as a comparison to the prior two as they are calculated from the inter-relationships of the clustered data with itself.

Each of the four metrics were tested on a set of 50 runs. The number of recovered clusters and the mode of the accuracy was taken for each set of 50 runs, Tables 1.1 and 1.2, respectively. Pure supervised metrics, where the ground truth is known, exhibited the best performance. The Accuracy metric achieved very high ratings, miss-matching only a few points. Rand performed well on the Iris dataset, less so on Wine. Interestingly, Rand found better results on the WBC dataset, than Accuracy. Neither the Silhouette nor the Dunn index performed well with any of the data. This is not surprising, due to the absence of ground truth in these indices and a lack of disparity in the index value for good and poor results.

Table 1.1. Mean and Variance of the Number of Clusters Recovered with a Given PSO Optimization Metric over 50 Runs

Clusters Recovered	PSO Optimization Metric			
	Accuracy	Rand	Silhouette	Dunn
Iris (3)	3±0	3.36±0.7494	2±0	2±0
Wine (3)	3±0	3.6±0.6061	2.24±0.4764	2±0
WBC (2)	2.86±0.3505	3.04±0.4020	2.02±0.1414	2±0

Table 1.2. Mode Accuracy of a Given PSO Optimization Metric over 50 Runs

Mode Accuracy per Metric	PSO Optimization Metric			
	Accuracy	Rand	Silhouette	Dunn
Iris	0.9667	0.9667	0.6667	0.6667
Wine	0.9775	0.7191	0.3371	0.3315
WBC	0.9048	0.9356	0.6706	0.6706

The PSO-ART implementation was them compared with generic Fuzzy ART and Fuzzy ARTMAP (Table 1.3). High performing vigilance values were chosen for each dataset. PSO-ART outperformed Fuzzy ART and Fuzzy ARTMAP in all instances, except with the WBC dataset. While PSO-ART found better results than Fuzzy ART, it did not outperform Fuzzy ARTMAP.

Table 1.3. Mean Accuracy Comparing Fuzzy ART, Fuzzy ARTMAP, and PSO-ART over 50 Runs

Mean Accuracy	Fuzzy ART	Fuzzy ARTMAP	PSO-ART
Iris	0.9333	0.9533	0.9663
Wine	0.9213	0.7191	0.9685
WBC	0.8199	0.9224	0.8805

1.5. CONCLUSION

Implementing per-cluster vigilance thresholds in ART has the potential to be of value for pattern recognition and discovery. Optimizing for vigilance allows each cluster to better represent its data. It also allows some clusters to be pushed away if their existence is not optimal. Both the Silhouette and the Dunn indices had the disadvantage of not having a high disparity in the range of values they can take. The lacking value disparity led to category abatement, or early stopping.

Adaptive Resonance Theory produces easy to understand clusters. It can be seen how much each cluster category fits an arbitrary feature of the data. With a vigilance threshold for each category, it can be seen how well a pattern must match a category for it to be considered a member.

This was not an exhaustive search of validation indices on which the Particle Swarm could optimize. Several indexes were, however, identified as candidate metrics. Current results show much better performance for external criteria as opposed to internal criteria. A good internal criterion would add useful autonomy to the ART implementation.

REFERENCES

[1] R Xu, G.C. Anagnostopoulos, D Wunsch. Multiclass Cancer Classification Using Semisupervised Ellipsoid ARTMAP and Particle Swarm Optimization with Gene Expression Data. IEEE.ACM Transactions on Computational Biology and Bioinformatics, vol 4, issue 1, pp 65-77, 2007.

[2] J Fonseca Antunes, N de Souza Araujo, C Minussi. Multinodal Load Forecasting using an ART-ARTMAP-fuzzy Neural Network and PSO strategy. IEEE Grenoble PowerTech, pp 1-6, 2013.

[3] R Xu, J Xu, D Wunsch. Clustering with Differential Evolution Particle Swarm Optimization. IEEE Congress on Evolutionary Computation, pp 1-8, 2010.

[4] L Meng, A H Tan, D Wunsch. Vigilance Adaptation in Adaptive Resonance Theory. IEEE International Joint Conference on Neural Networks, pp 1-7, 2013.

[5] J Kennedy, R Eberhart. Particle Swarm Optimization. IEEE International Conference on Neural Networks, vol 4, pp 1942-1948, 1995.

[6] F Li, J Zhan. Fuzzy adapting vigilance parameter of ART-II neural net. IEEE International Conference on Neural Networks, vol 3, pp1680-1685, 1994.

[7] P Werbos. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. PhD thesis, Harvard University, 1974.

[8] B Widrow, M Hoff. Adaptive switching circuits. IRE WESCON Convention Record, part 4, pp96-104.

[9] F Rosenblatt. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. 1962.

[10] Machine Learning Repository. http://archive.ics.uci.edu/ml/.

[11] G Carpenter. Default ARTMAP. Proceedings of International Joint Conference on Neural Networks, vol 2, pp 1396-1401, 2003.

[12] G Carpenter, S Grossberg. The ART of Adaptive Pattern Recognition by a Self-organizing Neural Network. IEEE Computer, vol 21, part 3, pp 77-88, 1988.

[13] G Carpenter, S Grossberg, D Rosen. Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System. Neural Networks, vol 4, pp759-771, 1991.

[14] G B Huang, Q Y Zhu, C K Siew. Extreme Learning Machine: A New Learning Scheme of Feedforward Neural Networks. IEEE International Joint Conference on Neural Networks, vol 2, pp 985-990, 2004.

[15] S Grossberg. How does the brain build a cognitive code. Psychological Review, pp 1-51, 1980.

[16] S Grossberg. Competitive learning: From interactive activation to adaptive resonance. Cognitive Science, vol 11, pp 23-63, 1987.

[17] S C Hsu, C F Chien. Hybrid data mining approach for pattern extraction from wafer bin map to improve yield in semiconductor manufacturing. International Journal of Production Economics vol 107, pp88-103, 2007.

[18] L Liu, L Huang, M Lai, C Ma. Projective ART with buffers for the high dimensional space clustering and an application to discover stock associations. Neurocomputing, vol 72, pp1283-1295, 2009.

[19] S Mulder, D Wunsch. Million city traveling salesman problem solution by divide and conquer clustering with adaptive resonance neural networks. Neural Networks, vol 16, pp827-832, 2003.

[20] A H Tan. Cascade ARTMAP: integrating neural computation and symbolic knowledge processing. IEEE Transactions on Neural Networks, vol 8, pp237-250, 1997.

II. TIME SERIES PREDICTION VIA TWO-STEP CLUSTERING

2.1. ABSTRACT

Linear and nonlinear models for time series analysis and prediction are well-established. Clustering methods have recently gained attention in this area. This paper explores a framework that can be used to cluster time series data. The range of values of a time series in clustered. Then the time series is clustered by data windows that flow into the initial set of value clusters. We can ensure with higher certainty that predictive temporal patterns are discovered across the whole range of values.

2.2. INTRODUCTION

2.2.1. Linear and Nonlinear Methods. Time series analysis and forecasting are each useful in a variety of scientific and engineering applications (e.g., weather forecasting, control, signal processing, and finance). The various types of models for analyzing and forecasting time series are linear models, nonlinear models, and clustering models.

Linear models (e.g., the moving average model [MA], the auto-regressive model [AR], and the auto-regressive moving average model [ARMA]) are popular for their well-defined statistical properties [9]. Linear models can break down when the time series has either a wide band spectrum or unknown seasonal components [8].

Nonlinear models (e.g., artificial neural networks) greatly extend the capacity to learn complex functions. Artificial neural networks allow for the distortion of the input space into a feature space that can be separated linearly [12]. The use of neural networks in time dependent domains requires the determination of time lags to be used in the

neural architecture. Although neural networks can be quite powerful, careful design decisions must be made that are not always intuitive.

Clustering methods, a subset of nonlinear models, are designed to uncover hidden structures in data. A time series already possesses a structure [9] (the temporal dependence) in addition to anything discovered analytically. Clustering methods should be able to discover temporal patterns that have predictive power.

2.2.2. Fuzzy ART. Adaptive Resonance Theory (ART) is an unsupervised learning theory. ART is capable of learning arbitrary data vectors in a stable and self-organizing way that overcomes the stability-plasticity dilemma [13-17]. A variant called Fuzzy ART [15] will be referred to for the remainder of this discussion.

Fuzzy ART is comprised of an input layer and a category layer. All input patterns are normalized between [0,1]. The weight matrix (w_j) acts as a category template. A category choice is calculated for each category against the input pattern:

$$T_j = \frac{|x \wedge w_j|}{\alpha + |w_j|} \tag{2.1}$$

where $\wedge$ is the fuzzy AND operator defined by

$$(x \wedge y)_i = \min(x_i, y_i) \tag{2.2}$$

and α is a choice parameter that is used to break ties.

In a winner-take-all fashion, the category with the largest T_j is chosen. A category match is calculated after a category choice is made, by comparing the winning node to the vigilance threshold:

$$\rho \leq \frac{|x \wedge w_j|}{|x|} \tag{2.3}$$

This determines the "goodness" of the match.

If the input pattern is classified as a match, that pattern is mapped to the selected node. The node is turned off via a reset mechanism if the node does not match, and a new competition in the category layer takes place. As each input pattern is matched to a category, the cluster mapping is build.

2.2.3. K-Means. The K-means algorithm [5] attempts to group n observations into k clusters. Optimal partitions are formed when the sum of squares error from each observation to its nearest centroid mean is minimized. Each centroid represents each of the k clusters.

K-means is easy to implement. Unfortunately, it can produce misleading results [6,7]. The most basic formulation is as [5,10]:

1. Initialize k partitions in a d-dimensional feature space
2. Assign each of the n observations to the nearest Partition (P_l) that has the smallest sum of squares to its centroid mean (m_l). For example,

$$x_j \in P_l, if \left\| x_j - m_l \right\|^2 < \left\| x_j - m_i \right\|^2 \forall j \in [1,n]; i \neq l; i,l \in [1,k] \quad (2.4)$$

3. Update the centroid means to reflect the observation's new partitions

$$m_i = \frac{1}{N_{P_i}} \sum_{x_j \in P_j} x_j \quad (2.5)$$

4. Repeat steps 2 and 3 until either a minimum threshold of iterations has transpired or no change occurs in the partition's make-up.

2.2.4. Two-Step Clustering. Preprocessing is an important step in data analysis. In this two-step clustering methodology, clustering serves as a step in preprocessing. The time series is clustered first by value (Fig. 2.1). This partitions the time series into value

bins which, essentially, performs vector quantization. The time series is then partitioned into n-step overlapping contrails (i.e., t(1:n), t(2:n+1), and so forth). These contrails are distributed among the value bins by their next value, t+1. Each group of contrails is then clustered to build prototype shapes that flow into each value bin (Fig. 2.2). These prototypes are created by averaging all of the contrails in that cluster. The cluster prototypes are finally compared against test data for t+1 predictions. The matching prototypes are chosen, and the corresponding target values are compared to the test data's target.

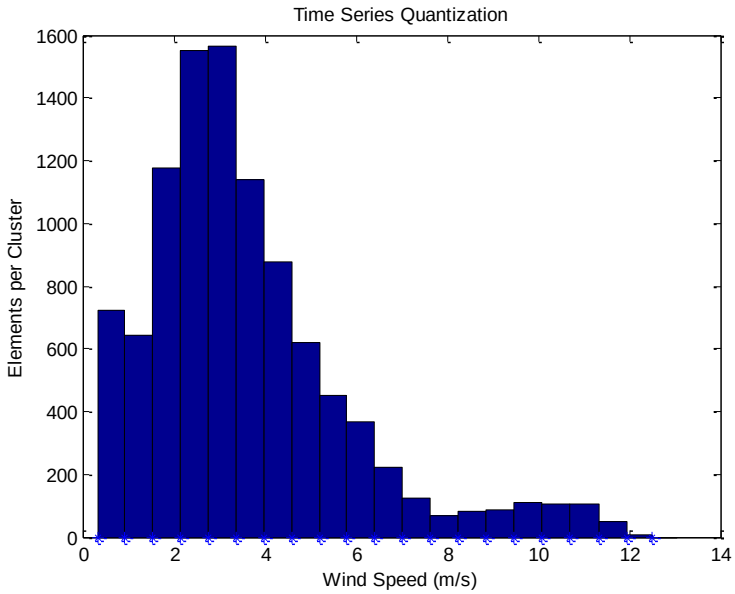

Figure 2.1. Wind Speed Time Series Quantized by Value

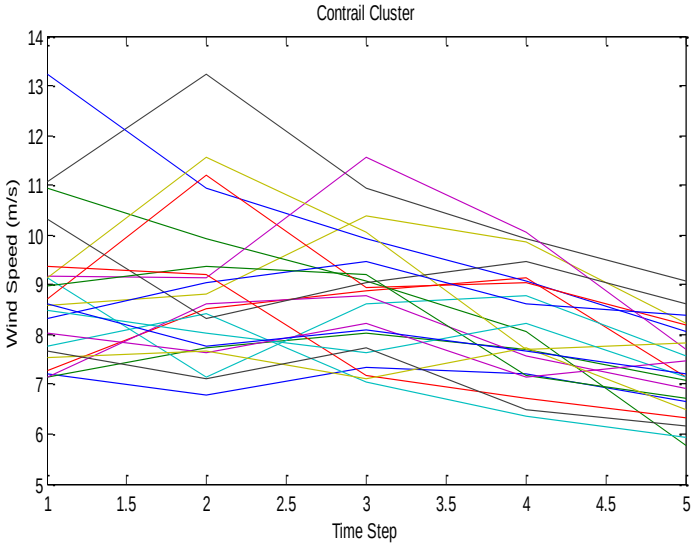

Figure 2.2. Wind Speed Contrail Cluster

This two-step clustering methodology acts as a framework for clustering time series. Different clustering methods can be interchanged for both target value clustering and contrail clustering.

2.3. DATA, EXPERIMENTS, AND RESULTS

Two datasets were used to test the utility of the proposed clustering framework. The first data set was taken from the National Renewable Energy Lab's (NREL) M2 Tower in Boulder, Colorado [11]. This data contained wind speed that has been recorded every 60 seconds. The training and testing data was collected from April 7, 2014 - April 13, 2014 and April 7, 2013 - April 13, 2013, respectively. The Mackey Glass equation was also used as its chaotic dynamics are of interest [19]. The Mackey Glass equation can be described as follows:

$$\frac{dx}{dt} = \beta * \frac{x(t-\tau)}{1+x(t-\tau)^n} - \gamma * x(t) \tag{2.6}$$

where β is equal to 0.2, γ is equal to 0.1, and n is equal to 10. Thirty thousand time steps were generated. The first 10,000 were used for initialization, the second 10,000 were used for training, and the third 10,000 were used for testing. A five-point moving average was used to smooth the training data.

Two clustering steps were included in the framework, and two cluster algorithms were chosen: K-means and Fuzzy ART. A total of four combinations were possible. Each combination was tested over 50 runs, see Table 2.1. All four methods performed comparably.

Table 2.1. Mean and Standard Deviation of the MSE of Time Series Predictions based on 50 Runs

Data Set	FuzzyART-FuzzyART	FuzzyART-kMeans	kMeans-FuzzyART	kMeans-kMeans
Wind Data	0.4318 ±0.0000	0.4296 ±0.0015	0.4240 ±0.0007	0.4204 ±0.0020
Mackey-Glass	1.354e-4 ±0.0000	1.1067e-4 ±0.0431e-4	1.3305e-4 ±0.0271e-4	1.1004e-4 ±0.0524e-4

With 60 prediction prototypes

A comparison was made with the individual algorithms that the two-step methods are comprised of (Table 2.2). All formulations of the two-step framework generated approximately 60 prototype vectors for prediction purposes. Compared to either k-means

clustering or fuzzy ART used alone, the performance of the two-step methods was better. The individual methods were tested up to approximately 500 partitions to get the best performance for comparison. The two-step methods were better with an order of magnitude less predictor prototypes.

Table 2.2. Comparison between Individual Methods' Mean and Standard Deviation of the MSE based on 50 Runs

Data Set	K-means*	Fuzzy ART**
Wind Data	0.7334 ±0.0537	0.5744 ±0.0000
Mackey Glass	5.760e-4 ±0.0604e-4	5.380e-4 ±0.0000

* K-means set to 500 partitions

** Fuzzy ART partitioned into 512 clusters

In Table 2.3, two sample t-Tests were performed to check if each of the two-step methods was better than the individual methods. All two-step formulations showed a significant performance difference with the p-value significant in all cases.

Table 2.3. Comparison of Two-Step Methods with Individual Methods using a t-Test

Data Set	Method	Vs	t-Test	
			p-value	Significance
Wind Data	kMeans-kMeans	kMeans	$<10^{-63}$	Extremely Significant
		Fuzzy ART	$<10^{-172}$	Extremely Significant
	kMeans-fuzzyART	kMeans	$<10^{-63}$	Extremely Significant
		Fuzzy ART	$<10^{-213}$	Extremely Significant
	fuzzyART-kMeans	kMeans	$<10^{-62}$	Extremely Significant
		Fuzzy ART	$<10^{-181}$	Extremely Significant
	fuzzyART-fuzzyART	kMeans	$<10^{-62}$	Extremely Significant
		Fuzzy ART	0	Extremely Significant
Mackey Glass	kMeans-kMeans	kMeans	$<10^{-160}$	Extremely Significant
		Fuzzy ART	$<10^{-175}$	Extremely Significant
	kMeans-fuzzyART	kMeans	$<10^{-166}$	Extremely Significant
		Fuzzy ART	$<10^{-200}$	Extremely Significant
	fuzzyART-kMeans	kMeans	$<10^{-163}$	Extremely Significant
		Fuzzy ART	$<10^{-183}$	Extremely Significant

Table 2.3. Comparison of Two-Step Methods with Individual Methods using a t-Test (cont.)

Data Set	Method	Vs	t-Test	
			p-value	Significance
Mackey Glass	fuzzyART-fuzzyART	kMeans	$<10^{-170}$	Extremely Significant
		Fuzzy ART	0	Extremely Significant

2.4. CONCLUSION

The two-step clustering framework applied to time series data exhibited promising results over individual methods, as confirmed by t-Test results. Quantization of the time series helps ensure that prototypes can be generated across the entire range of data. K-means and Fuzzy ART were applied together and separately in all possible combinations. The performance of each two-step formulation produced results that were relatively similar, and all were superior to the corresponding techniques in isolation.

REFERENCES

[1] J Zakaria, A Mueen, E Keogh. Clustering Timer Series Using Unsupervised-Shapelets. IEEE 12th International Conference on Data Mining, pp 785-794, 2012.

[2] A.B. Geva. Non-Stationary Time-series Prediction using Fuzzy Clustering. 18th International Conference of the North American Fuzzy Information Processing Society, pp 413-417, 1999.

[3] F Liu, P Du, F Weng, J Qu. Use Clustering to Improve Neural Networks in Financial Time Series Prediction. Third International Conference on Natural Computation, vol 2, pp 89-93, 2007.

[4] B Chandra, M Gupta, M.P. Gupta. A Multivariate Time Series Clustering Approach for Crime Trends Prediction. IEEE International Conference on Systems, Man, and Cybernetics, pp 892-896, 2008.

[5] J MacQueen. Some methods for classifiation and analysis of multivariate observations. Proceedings of the 5th Berkley Symposium on Mathematical Statistics and Probability, Statistics, vol 1, Univercity of California Press, 1967.

[6] J Wu, H Xiong, J Chen, W Zhou. A Generalization of Proximity Functions for K-means. Proceedings of the 2007 IEEE International Conference on Data Mining, 2007.

[7] H Xiong, J Wu, J Chen. K-means Clustering versus Validation Measures: A data distribution perspective. Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining, pp 779-784, 2007.

[8] J.D. Hamilton. Time Series Analysis. Princeton University Press, 1994.

[9] P Brockwell, R Davis, Introduction to Time Series and Forecasting, Second Edition. Springer, 2010.

[10] R Xu, D Wunsch. Clustering. Wiley, 2009.

[11] National Renewable Research Lab's M2 Tower, https://www.nrel.gov/midc/nwtc_m2/.

[12] P Werbos. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. PhD thesis, Harvard University, 1974.

[13] G Carpenter. Default ARTMAP. Proceedings of International Joint Conference on Neural Networks, vol 2, pp 1396-1401, 2003.

[14] G Carpenter, S Grossberg. The ART of Adaptive Pattern Recognition by a Self-organizing Neural Network. IEEE Computer, vol 21, part 3, pp 77-88, 1988.

[15] G Carpenter, S Grossberg, D Rosen. Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System. Neural Networks, vol 4, pp759-771, 1991.

[16] S Grossberg. How does the brain build a cognitive code. Psychological Review, pp 1-51, 1980.

[17] S Grossberg. Competitive learning: From interactive activation to adaptive resonance. Cognitive Science, vol 11, pp 23-63, 1987.

[18] F Rosenblatt. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. 1962.

[19] M.C. Mackey, L Glass. Oscillation and Chaos in Physiological Control Systems. Science, vol 197, pp 287-289, 1997.

SECTION

2. CONCLUSION

2.1. CLOSING THOUGHTS

In this research, ART was examined for its extensibility and applications. ART is limited by a single vigilance value that controls the performance of the implementation. By assigning a vigilance value to each cluster and optimizing them with a PSO implementation, this extension outperformed Fuzzy ART on three datasets and Fuzzy ARTMAP on two datasets, out of three total datasets. ART and K-Means were examined as a means of performing vector quantization. This clustering quantization boosted prediction results when applied to time series.

VITA

Clayton Parker Smith was born August 28th, 1989. He received his BS in Computer Engineering from the Missouri University of Science and Technology in May 2013. He has been enrolled in Graduate School at Missouri University of Science and Technology since January 2013. Since graduate enrollment, he has worked as a graduate research assistant in the Applied Computational Intelligence Laboratory in the Department of Electrical and Computer Engineering. His current degree is a MS in Computer Engineering to be awarded in May 2015.