

Research Statement

Overview of Current and Long Term Research Goals

In my current research, I am studying the training of fuzzy measures by classic and stochastic optimization in information fusion. In my dissertation research, I proposed several novel algorithms for the task. These algorithms range from a classic gradient descent method to stochastic methods. Currently, I am integrating these novel algorithms into cost functions for non-convex clustering. This will enhance the decision level fusion properties of the clustering algorithms. Furthermore, I am looking at receiver operating characteristic curve optimization to improve the fusion properties of my dissertation algorithms.

In my long term research, I plan to study of the use of non-parametric distributions for problems in database and data mining. I believe that non-parametric distributions are the best approximations for the decision functions on these types of problems. I also want to explore the properties of the space of classification algorithms to understand their mutual relations. This can allow us to define new functional metrics to compare them. Finally, several researchers have pointed out that the use of generic classification algorithms in problems like gene classification is inefficient. Therefore, it is necessary to find better mathematical models for these types of problems.

Dissertation Research

In practical applications of pattern classification, multiple algorithms are often developed for the same classification problem. Each algorithm produces confidence values by which each new sample may be classified. We would like to aggregate these confidence values to produce the best possible confidence for the given sample. This can be seen as a particular instance of what is called information fusion.

Besides learning parameters of aggregation operators to assign the best confidence for a given sample, we would also like the aggregation operators to use a subset of the algorithm confidences and achieve the same level of performance as the entire set of confidences. Using a subset of the algorithms implies lower cost for applications.

Choquet integrals are nonlinear operators based on fuzzy measures that can represent a wide variety of aggregation operators. These integrals have been proposed for use in pattern classification and information fusion by several researchers. Previous research has demonstrated the utility of Choquet integrals for information fu-

sion compared to other methods such as neural networks and Bayesian approaches.

Under several grants from the Army and the National Science Foundation, my dissertation research developed algorithms for learning of fuzzy measures for decision level fusion in the landmine detection problem.

One novel result of my dissertation research is that the fuzzy measures learned can be very sensitive to the choice of desired outputs. In response to this problem, I developed an alternative training methodology based on Minimum Classification Error (MCE) training that does not require the use of desired outputs. I demonstrated that better performance can be achieved using this training methodology. The application and demonstration of the utility of MCE training for learning fuzzy measures is novel, and the methodology has been applied to several problems in landmine detection. A problem with this method is that it depends on a constrained type of fuzzy measure, the Sugeno measure.

There was a need for additional approaches to learning unconstrained fuzzy measures that are more computationally attractive and provide robust performance. I proposed an approach to learning unconstrained fuzzy measures that relies on Markov Chain Monte Carlo sampling methods. The use of such approaches for learning measures for Choquet integral fusion was completely novel. In addition, I proposed the inclusion of the Bayesian approach of imposing sparsity promoting prior distributions on the measure parameters during sampling as a way of selecting subsets of the algorithms for inclusion in the aggregation. This approach was completely new for learning fuzzy measures.

Several algorithms were developed by using the new stochastic approach. First, a collection of algorithms based on Gaussian and Laplacian learning functions when the desired outputs are well known. Second, a set of algorithms based on the Logistic distribution learning function to train one or two measure models when the outputs are unknown.

Current Research

Currently, I am involved on integrating the previous methods into cost functions for non-convex clustering for the project “Optimized Multi-algorithm Systems for Detecting Explosive Objects Using Robust Clustering and Choquet Integration,” sponsored by the National Science Foundation. The members of this project are considering cost functions that cluster data into non-convex sets. It has been observed that landmine data classes tend to overlap. Therefore, it makes sense to use cost functions that simulate this behavior.

In this project, I am proposing to move the cost functions from the classic opti-

mization framework, gradient descent and Laplace multipliers, to a Bayesian framework to apply my dissertation methods. This will allow me to use the non-convex properties of the generated clusters to improve the fusion of the clustering function.

Additionally, I am looking at the optimization of the Receiver Operating Characteristic (ROC) curves by probabilistic methodologies. I want to integrate these ROC optimizations into my dissertation algorithms to improve confidence fusion and false alarm rate.

Finally, I am looking at the use of k-additive fuzzy measures to increase the flexibility of my dissertation algorithms to learn the measure parameters in the fusion problem.

Short Term Research Goals

In my short term research, I plan to implement the training of the Sugeno measure and other recursive fuzzy measures in the multinomial logistic regression methodology. This will reduce the complexity of learning fuzzy measure parameters by stochastic simulation to linear time. Using this new algorithm, I plan to study data sets with large number of features to explore dimensionality reduction. Examples of these types of data sets are in gene classification and context handwritten decision level fusion.

In addition, I am going to study the rates of convergence of my dissertation algorithms. Finding these rates of convergence can allow me to reduce their time complexity. This can make them more suitable to study data sets with dozens of possible features.

Further, I noticed that the lattice of the power set under the fuzzy measure constraints can be seen as a Bayesian network. This has given me new ideas on how to train the fuzzy measures by methods used in Bayesian networks.

Finally, many dissimilarity measures are not efficient enough to improve the separation of classes in the fusion problem. Therefore, it is necessary to study new dissimilarity operators for fusion. For this, I am considering operators based on manifold metrics and information theory pseudo metrics.

Long Term General Research Goals

For my long term research, I am interested on the study of more general learning functions derived from stochastic processes and non-parametric distributions. In many problems, like the ones in databases and data mining, it is not possible to assign a closed form for the distribution of the data. Instead, it is only possible to

generate non-parametric distributions. Therefore, it is necessary to study effective ways to generate and manipulate these non-parametric distributions. Once these distributions are generated, we can use them to study:

- Probabilistic properties of each data class.
- The causality relations between the different data classes.
- Algorithms that identify and generate new causal relations as data evolves in time.
- etc.

Classic examples of these non-parametric distributions are the ones generated by Dirichlet processes.

Additionally, I am interested in studying the general properties of the classification algorithms. Something that has been pointed out several times is that we have many classification algorithms, but no good ways to compare them. Therefore, it is necessary to look at more general theories of the space of classification algorithms. Consequently, we need to ask ourselves the following questions:

- Does a spanning base exist for this space?
- Do we have equivalence relations that allow us to partition the space?
- What are the commonalities between the algorithms in a single partition?
- etc.

Answering these questions will allow us to have a deeper understanding of the space of classification algorithms.

Finally, an important problem is the mathematical modeling of systems where few samples exist and each of them contains thousands of features. It has been pointed out by several researchers that under these conditions data driven classification algorithms do not perform well. Therefore, what we may need is not a new classification algorithm, but new mathematical models for these types of problems. This can have applications in gene classification or stream data classification.