Q-Learning Multi-Objective Sequential Optimal Sensor Parameter Weights

Raquel COHEN
Mark RAHMES
Kevin FOX
George LEMIEUX
Harris Corporation
Melbourne, Florida USA

ABSTRACT

The goal of our solution is to deliver trustworthy decision making analysis tools which evaluate situations and potential impacts of such decisions through acquired information and add efficiency for continuing mission operations and analyst information. We discuss the use of cooperation in modeling and simulation and show quantitative results for design choices to resource allocation. The key contribution of our paper is to combine remote sensing decision making with Nash Equilibrium for sensor parameter weighting optimization. By calculating all Nash Equilibrium possibilities per period, optimization of sensor allocation is achieved for overall higher system efficiency. Our tool provides insight into what are the most important or optimal weights for sensor parameters and can be used to efficiently tune those weights.

Keywords: Game Theory, Resource Management, Modeling and Simulation, Augmented Decision Making, Q-Learning.

1. DECISION MAKING APPROACH

Human decision making activities performed with data from disparate sources is difficult and a highly time consuming activity in near real time or on demand modes. Human cognition and knowledge base within the decision making process must also be considered as an important factor. There are additional needs for increased information analysis capabilities demonstrating more accurate decisions, planning factors, resource allocation, risk management, and information analysis in a near real time, visually oriented manner with fewer analysts and mission planners.

A major goal within industry and others is to push forward an open architecture framework in order to: inject and fuse data and information from a multitude of sources, contain collaborative environments, provide increased visualization of information (immersion), improve decision making performance in analysis and mission planning, and increase pattern recognition among disparate data sets in order to effectively analyze information.

One way to address the decision making process from the human approach for analysts and mission planners is the use of serious games or simulated environments. Serious games can provide simulated virtual learning venues for mitigation of selected biases found within human decision making process [5]. Training and simulations in virtual environments can also allow for immersive simulations and training of real world scenarios thus potentially increasing performance within human decision making process.

The sampling of continuous Earth-observation data significantly simplifies the problem of sensor allocation as shown in Fig 1. We allow our system to allocate a sensor resource at a given epoch. Of course, the time period can be modified per user specification. The feedback loop accounts for last time the Area of Interest (AOI) was collected. We combine the Nash Equilibrium's (NEs) from each dimension by a weighted sum. This methodology gives an analyst sufficient control over model. Our research to date has shown that it is more efficient to combine each NE from each dimension rather than combine all reward matrices and then calculate NE. This method also allows for more control and weighting of value of each dimension or category. Additionally, equalizing units from each dimension is important [10].

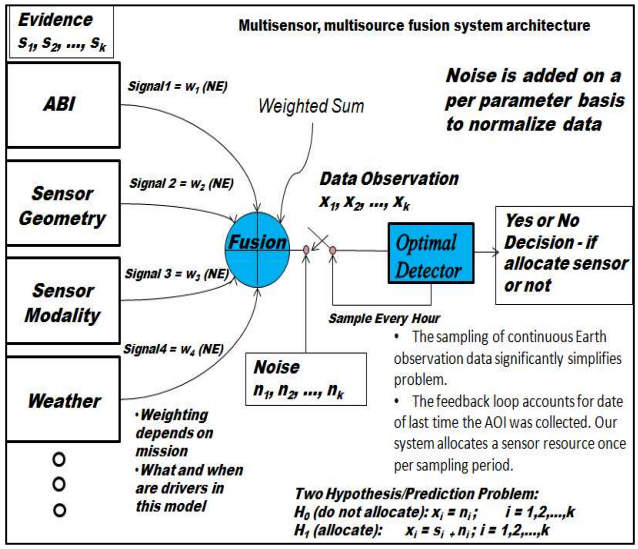

Fig 1. Example Fusion and Decision-making Framework

Current situational awareness efforts seek to incorporate not only geospatial features and structures, but also the human element, especially in urban settings. Development of tools for more rapid refinement of flexible plans is required for adapting to a changing operational environment. Our work can enable a methodical approach to intelligent planning and reaction including interaction of variables, parameters and attributes by the user resulting in updated probabilities.

2. INFORMATION FUSION

Our solution uses a modified Dempster's Rule of combining evidence with Nash Equilibrium (NE). The goal is to gather evidence from several AOIs and determine how to allocate limited sensor resources to maximize data collection. Sensor geometry, along with sensor resolution, determines when and what can be sensed. Activity Based Intelligence (ABI) and weather are critical for where to image. This concept is shown in Fig 2. There will be several choices that meet all the evidence criteria.

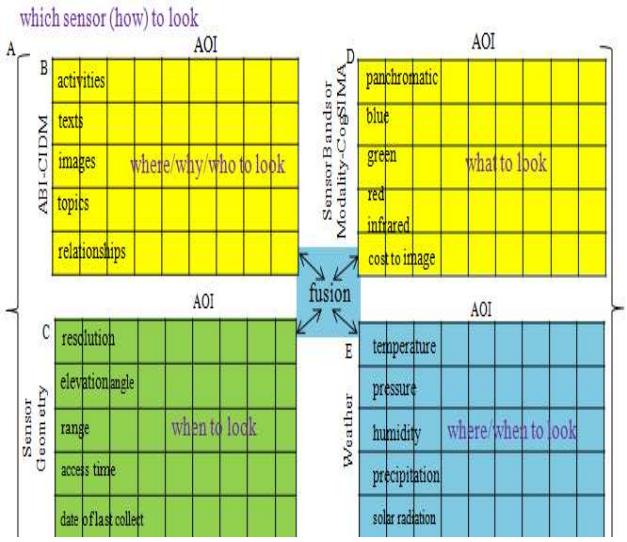

Fig 2. Multi-Dimensional Reward Matrices

Dempster-Shafer evidential reasoning for multi-sensor fusion allows each sensor to contribute information at its own level of detail. Dempster reasoning is an extension of Bayesian approach that makes explicit any lack of information concerning probability. All propositions for which there is no information are assigned an equal a priori probability [1].

Our example can be modeled with sensors (columns in reward matrix), parameters (rows in reward matrix), and AOIs (third dimension in reward matrix). The Dempster Rule, along with NE, provides a framework for combining evidence from all dimensions. The parameter measurements of system are stored in the reward matrix and used in the mathematical model to determine optimal sensor for an AOI. The sensors and AOI are players in this game. In our example, dimensions correspond to the letters A, B, C, D and E. The A values are intersection of all reward matrices from each category and determine which sensor to allocate.

Our multi-dimensional solution populates a reward matrix for each parameter in near real time through powerful game theory analysis. Once data accuracy is proven through sensitivity analysis, the information can either be used as training data, or populated into a reward matrix for resource allocation and adversarial planning utilizing game theory concepts such as in a competitive or cooperative game model. Much of the current focus is on human geography and terrain, as well as population-based sentiment analysis [10]. Parameters for sensor geometry dimension include: pixel resolution (cm); elevation (deg); range (km); access duration (sec); and most recent collection date (days old).

3. GAME THEORY

Game theory is the study of strategic decision-making and mathematical modeling of conflict and cooperation between intelligent, rational decision-makers, and is often thought of as an interactive decision theory. It has been applied to economics, political science, psychology, logic, biology and other complex issues. Modern game theory began with the idea of the existence of mixed-strategy equilibrium in two-person zero-sum games, applied to economics. Later, this evolved to provide a theory of expected utility, which allowed mathematicians and economists to treat decision-making with uncertainty. The notion of probabilistic predictions utilizing game theory is critical to many decision-making applications because optimizing user experience requires being able to compute expected utilities of mutually exclusive data.

Maximin equilibrium often is the strategy and is called the Nash theory application of zero or constant sum strategy game [7]. Game theory considers the effect of a player's decision on other decision-makers. In many situations, the opponents know the strategy that they are following and what actions are available. The Nash threshold can be used to determine if the player is on the blue or red team. For example, if a reward matrix exists, then the equilibrium point is the one where the reward is the smallest value in its row and the largest number in its column [13].

$$max_{all\ rows}(row\ min) = min_{all\ columns}(column\ max) \qquad (1)$$

This left half of (1) presents the basic applied theory to decision-making of our model under uncertainty. For a possible action, one consideration is to choose the "best" worst outcome. The maximin criterion suggests that the decision-maker should choose the alternative, which maximizes the minimum payoff he can get. This pessimistic approach implies that the decision-maker should expect the worst to happen. The maximin criterion is concerned with making worst possible outcome as pleasant as possible [13].

The right half of (1) represents minimax regret criterion, which uses concept of opportunity cost to arrive at a decision. The regret of an outcome is the difference between the value of that outcome and the maximum value of all the possible outcomes. For any action and state, there is opportunity of loss or regret. The decision-maker should choose the alternative that minimizes the maximum regret he could suffer [13].

Equation (2) is a translation of a reward matrix to a linear program that can be solved mathematically. We calculate the NE for each reward matrix [7]. Linear programming is useful for solving game theory problems and finding optimal strategies. We can define:

x_1 = probability that blue player chooses Worldview 1
x_2 = probability that blue player chooses Worldview 2
x_3 = probability that blue player chooses Worldview 3
x_4 = probability that blue player chooses Ikonos
x_5 = probability that blue player chooses QuickBird2

As an example, using reward matrix we show linear program solution for constant sum game as follows:

$$maximize\ v\ s.t. \qquad (2)$$
$$v - a_{11}x_1 - a_{21}x_2 - a_{31}x_3 - a_{41}x_4 - a_{51}x_5 \leq 0$$
$$v - b_{12}x_1 - b_{22}x_2 - b_{32}x_3 - b_{42}x_4 - b_{52}x_5 \leq 0$$
$$v - c_{13}x_1 - c_{23}x_2 - c_{33}x_3 - c_{43}x_4 - c_{53}x_5 \leq 0$$

$v - d_{14}x_1 - d_{24}x_2 - d_{34}x_3 - d_{44}x_4 - d_{54}x_5 \leq 0$
$v - e_{15}x_1 - e_{25}x_2 - e_{35}x_3 - e_{45}x_4 - e_{55}x_5 \leq 0$
$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 1$
$x_1, x_2, x_3, x_4, x_5 \geq 0$

The initial solution for optimal player's mixed strategy in terms of probabilities: x = (x1, x2, x3, x4, x5).

4. SENSOR RESOURCE ALLOCATION

In a fixed-prioritization approach, sensor resources are applied to tasks starting from the highest priority task and progressing to lower priority tasks as available resources allow. Each task will generally contain some constraints on allowable task execution times or rates depending on the type of task to be performed. Typically, tasks are pre-scheduled by a sensor's scheduling function in fixed time intervals or scheduling intervals (SIs) prior to execution. Within the SI, tasks are scheduled according to priority until either SI is filled or task list is exhausted. If task list is exhausted, then it is assumed that the sensor's available resources were sufficient to service all required tasks. However, if the SIs are being filled before completing the list of tasks in the queue to be scheduled and task time constraints are not being met, then the sensor is considered to be resource constrained. In a fixed-prioritization scheme there is no guarantee that lower priority tasks will be serviced at all, or if they are, that they will meet time scheduling constraints. Extending the length of the SI will result in more tasks being scheduled per interval, but may jeopardize meeting scheduling constraints on all tasks, including higher priority tasks [3].

We want to maximize overall system performance. Current model allocates one sensor per AOI. Output is allocation decisions. Our Optimal Detector uses NE Scoring. A hit is correct AOIs chosen. A miss is correct AOIs not chosen. A false alarm is incorrect AOIs chosen. A correct rejection is zero values in incorrect AOIs.

In a pure strategy game, the NE is the objective function that is the value of the game. Game theory serves as a framework for managing system inputs and outputs. NE provides a confidence value for a linear programming solution. Our algorithm applies linear programming to create a reward matrix. Determining which sensor to use involves calculating the maximum sum of all requests based on a reward matrix using a linear program.

In our example, there are several resource management stages including information needs, collection objectives, observables, tasks and plans. Resource management process seeks to decompose information needed to satisfy mission objectives into one or more tasks. The essence of resource management is uncertainty management [6]. Resource allocation problems in which limited resources must be allocated among several activities are often solved by linear programming. Operations Research is a branch of mathematics that studies decision-making to obtain best decision. Game theory can help determine optimal strategy [13].

5. PARAMETER MANAGEMENT

Some tools use "strategies" are measured in different units in the same reward matrix and can be problematic. Examples include use of manpower (count of people) mixed with propaganda (not necessarily units of people). If all strategies in a given decision model reward matrix are not in the same (equalized) units, then use of game theory and mini-max or maxi-min functions can provide misleading results. We can create purely dominant and incorrect solutions just due to relative size of unit measures. Our solution addresses this properly and uniformly for any decision model. We equalize all strategies (in a given decision model) to the same unit. This is a key point to the application of game strategies to a general class of decision problems. An adjustable "equalization" factor has the purpose to convert all strategy measures to the same unit (e.g., cost, time) and must be done for any decision model. The equalization factor for our solution is independent of additional (importance) weights that may be applied [9].

Using different weights for choices highlights the importance of an AOI, a sensor, or a parameter. A tool that can allow the user to dial and modify the weights of modeled parameters is important to model "what if" scenarios. Additionally, saving the weights to a file allows for peer review in order to check and validate decisions. Our approach is modeled, so that the process can be repeated to allow for new or higher-quality data/information to be inserted into the process to generate updated results [6].

In our example, it is straightforward to allocate one sensor to one AOI. However, the problem is not obvious for how to optimally determine parameter weights since there are N parameters per sensor. This is an underdetermined system. We can determine the weights of the sensor parameters through the use of Nash Equilibrium.

We use an interior-point algorithm, the primal-dual method, which must be feasible for convergence. The primal standard form (used to calculate optimal AOIs and Sensors) is:

$$minimize\ (f * x)\ s.t. \quad A * x = b,\ x \geq 0 \tag{3}$$

The dual problem (used to calculate optimal Parameters) is:

$$\boldsymbol{maximize}\ (\boldsymbol{b'} * \boldsymbol{y})\ \boldsymbol{s.t.} \quad \boldsymbol{A'} * \boldsymbol{y} + \boldsymbol{s} = \boldsymbol{f},\ \boldsymbol{s} \geq \boldsymbol{0} \tag{4}$$

Since we know the optimal sensor to allocate to a given AOI, we can find the associated column or parameter (column), given the row (sensor). Then we use the error to determine the optimal weights and importance of each parameter.

We create a triplet of weights for AOI, Sensor, and Parameters. This also makes it possible to use AOI with NE to determine parameter weights. This design is shown in Fig 3. The use of information from sequential time epochs allows additional insight into how fast a parameter weight can be learned relative to other reward volumes.

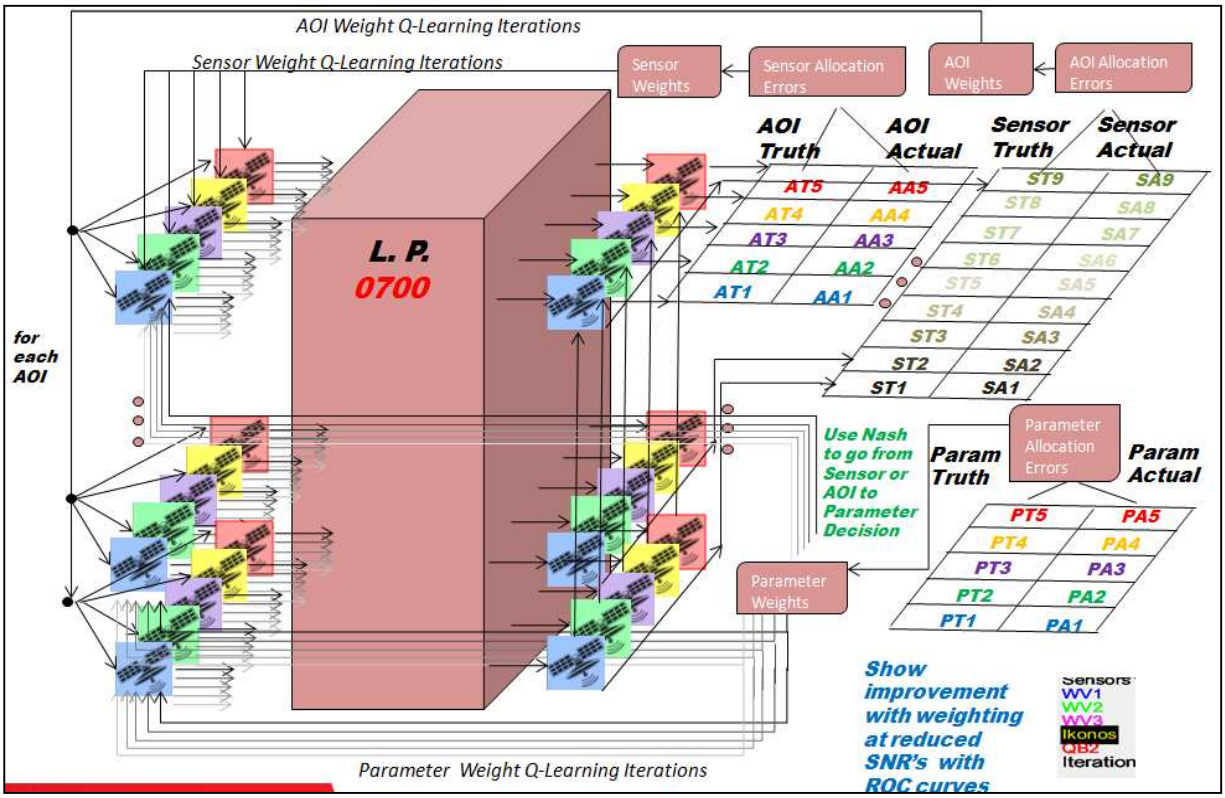

Fig 3. Q-Learning Block Diagram of Weight Triplets – AOI, Sensor, Parameters

We extend the 2D reward matrix to a 3D reward volume. In our case the players are AOI, Sensor, and Parameter. This new concept of reward volumes gives the decision maker an ability to correctly and automatically analyze multiple factors.

6. Q-LEARNING MODELING AND SIMULATION

Q-Learning can be used for function approximation which makes it possible to apply the algorithm to large problems, even when state space is continuous, and therefore infinitely large [4]. Our solution seeks to learn what the weights for each dimension should be to achieve optimal system performance. Our solution may also save money by offering a Pareto efficient, repeatable process for resource management.

Our system uses Q-Learning to assign optimal weights as a Markov decision process (MDP). Q-Learning is beneficial for determining weights for each dimension in our system and can give insight into relationships among objectives, improving the understanding of the problem. While each period is modeled as independent, four dimensions within a period are considered a dependent, sequential Markov function [8].

We use an optimal multi-objective Markov action-selection decision making function with Q-Learning. We currently consider input from activity based intelligence (ABI), sensor geometry, sensor modality, and weather. However our system is modular and flexible to handle any number of inputs. By optimizing weights from these inputs multiplied by the Nash Equilibrium (NE) values for each of the dimension possibilities per period, optimization of sensor allocation is achieved for overall higher system efficiency.

The Sensor Geometry determines the scheduling or when a sensor is available for a given area of interest (AOI). The activity based intelligence (ABI) section depends on the sensor geometry dimension. The ABI section answers where, why, and who to observe. The Sensor Modality dimension deals with the what. The weather dimension handles the where and when. These four dimensions with conflicting objectives depend on each other for an optimal solution of how to look or what sensor should be used for observation.

In our simulation, the truth data is the crisis value or priority determined by the daily executive requirements meeting. The values shown are examples for indicating the relative importance of the sensor type for a given AOI. In our concept of operations the weighted values can be derived based on the requirements as per daily executive requirements priorities. The executive daily meeting sets requirements for sensor modality to be used for an AOI. However the daily meeting does not assign a bird to an AOI. An algorithm then involves running a linear program for each sensor type and number until all assets are optimally tasked. When the action value weights are learned, the optimal policy can be constructed. Fig 4 shows the workflow to determine optimal AOI weights using Q-Learning with minimum mean square error (MMSE) calculation. The MMSE estimator is a common estimation method which minimizes the mean square error, in our case, between requirements and information in input dimensions.

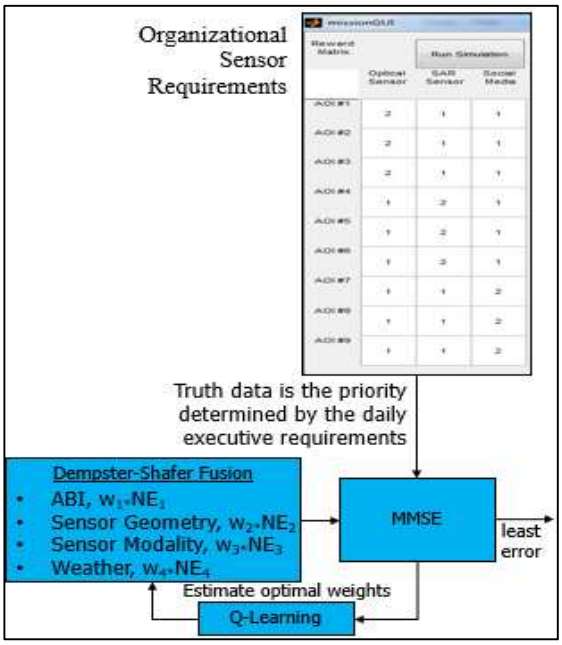

Fig 4. Q-Learning Weighting Optimization

When you use a mathematical model to describe reality you must make approximations. The world is more complicated than the kinds of optimization problems that we are able to solve. Linearity assumptions usually are significant approximations. Another important approximation comes because you cannot be sure of the data that you put into the model. Your knowledge of the relevant technology may be imprecise, forcing you to approximate values in a, b, or c in a linear equation. Moreover, information may change. Sensitivity analysis is a systematic study of how sensitive solutions are to changes in data. [2]

When such an action-value function is learned for weighting, the optimal policy can be constructed by simply selecting the action by combining the values from each AOI in each period using Dempster-Shafer Rule [11].In our example, we define the Q-Learning equation as:

$$newWeight = \frac{(1-errorA,B,C,D,E_{nom})*(NashEquilibriumValue)}{\sum_{vparams}(1-errorA,B,C,D,E_{nom})*(NashEquilibriumValue)} \quad (5)$$

where A, B, C, D, and E are parameters.

$$QLearnedWeight = oldWeight + learningRate * (newWeight - oldWeight) \quad (6)$$

The results of our simulation are shown in Fig 5. We normalized the parameter weights such that they add up to one. To begin with, we initialize all the parameter weights to one. Then we learn which parameters are most important for optimal settings. We set weights of AOIs and Sensors and tune the Parameters weights. Tuning the Parameter weights serves as a useful tool while providing insight to system. The innovation is to use the NE for solving Parameter weights given AOI and Sensor weights.

The learning rate determines to what extent newly acquired information will override old information. A factor of 0 will make the agent not learn anything, while a factor of 1 would make the agent consider only the most recent information. The discount factor determines importance of future rewards. A factor of 0 will make agent short-sighted by only considering current rewards, while a factor approaching 1 will strive for a long-term high reward. The initial condition for the estimate Optimal Weight can be set to the reciprocal of the number of dimensions or 0.25 which initially considers each of four dimensions as equally important.

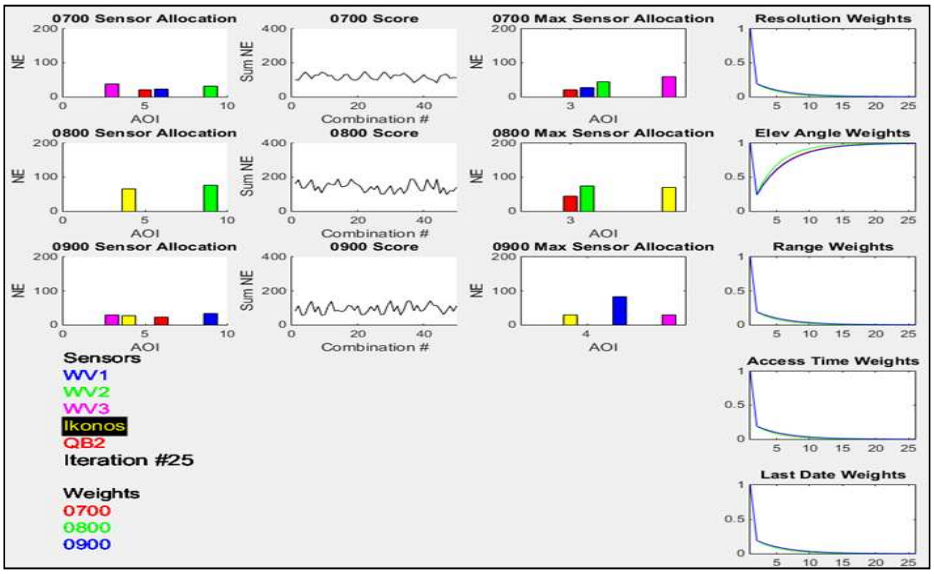

Fig 5. Iterative Q-Learning Simulation

7. ACCURACY ASSESSMENT

Classical decision theoretic scenario is that of an agent having to choose among a set of actions, consequences of which depend either on certain states of affairs about which agent is not completely informed, i.e., subjective uncertainty, or the result of some random processes that are independent, i.e., objective uncertainty [12].

The question of how do we know we have made best decision arises. Hard decisions such as using Boolean Logic to for Access times can

be used to help rule out some decisions since the sensor cannot see the AOI. Resolution determines if activity can be detected using NIIRS value or definition of activity of interest and if it will meet mission requirements. Soft decisions using game theory serve as a structure to develop models that better predict actual behavior. Simulation of system performance can be shown as a function of SNR. Values used in constructing hypothesis histograms are from a reward matrix. Detection theory allows for design and level of expectation of system performance with value of abnormal observations and measurements. Fig 6 shows our sensitivity analysis using several different signal-to-noise ratios (SNRs) under perfection conditions. The graph shows accuracy as a function of SNR. In our example we have added Gaussian noise to each parameter in the reward matrix. The SNR, d, is the distance between means on two hypotheses with a variance normalized to one. ROC Curves are shown for SNR = 12, 14, and 16. Histograms are shown for SNR=12.

ROC curve calculates the probability of detection at all thresholds. We calculated NE for 50 trials and all combinations of sensor allocations to all AOIs. The Pd is probability of deciding "signal present" given that it is present. The Pf is the probability of deciding "signal present" given noise alone. Our goal is to design a signal process that makes the "best" decision for sensor allocation. We evaluate performance of our algorithm and compare with sub-optimum approaches. We want to minimize probability of decision error and maximize the probability of a correct decision. ROC curve is a plot of Pd and Pf as a function of all possible threshold settings. We use the detectability index $d = \sqrt{Es/\sigma 2}$ where Es is the energy of the signal symbols. As the SNR increases, the ROC curve performance is improved.

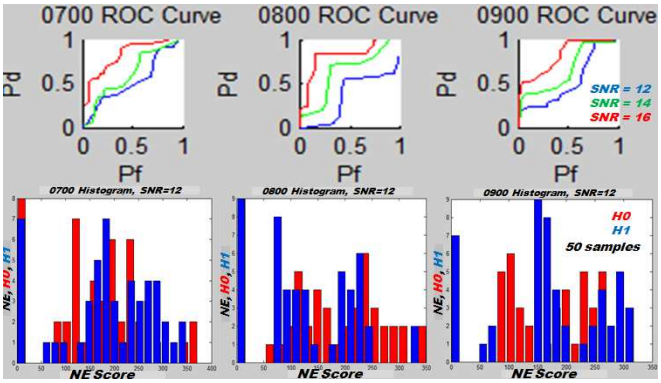

Fig 6. Iterative Q-Learning Simulation

8. CONCLUSIONS

The key contribution of our paper is to combine remote sensing decision making with Nash Equilibrium for sensor optimization. By calculating all Nash Equilibrium possibilities per sampling period, optimization of sensor allocation is achieved for overall higher system efficiency.

We have identified a novel mathematical application for sensor prioritization by collapsing multi-dimensional problems to use linear programming optimization. We calculate optimal strategies, resource allocation and increased likelihood of best decision available using game theory in a zero or constant sum game. The sampling of continuous Earth observation data significantly simplifies the problem.

Finally, we discussed a method for modeling asset management with limited resources for multiple sensor modality requirements. One solution is to run the Nash algorithms for each successive tasking request and then run a dynamic fair division water fill algorithm to ensure that each request is fair with respect to limited available assets. The motivation for fairness is used to ensure that not all sensor assets are dominated by one agent or player (region of interest). This is needed so as not to miss important events occurring around the world.

REFERENCES

[1] Abidi, M. A. "Fusion of multi-dimensional data using regularization." **Data Fusion in Robotics and Machine Intelligence** (1992): 415-455.

[2] Demers, A., Keshav, S., Shenker, S., "Analysis and simulation of a fair queueing algorithm". In **Proc. Of SIGCOMM**, pages 1–12, 1989.

[3] Geckle, W., Smoot, J., & Dockery, D. Modeling and Simulation of Sensor Task Assignment and Scheduling in CG (X), 2008.

[4] Hado van Hasselt. Reinforcement Learning in Continuous State and Action Spaces. **Reinforcement Learning: State of Art**, Springer, pages 207-251, 2012.

[5] IARPA's Sirius Program: http://www.iarpa.gov/Programs/ia/Sirius/presentations/Sirius_Overview.pdf.

[6] Liggins, Hall, Llinas, "Handbook of Multisensor Data Fusion, Theory, and Practice", 2nd Edition, 2009.

[7] Nash, John (1951) "Non-Cooperative Games" **The Annals of Mathematics** 54(2):286-295.

[8] Simone Parisi, Matteo Pirotta, Nicola Smacchia, Luca Bascetta, Marcello Restelli: Policy gradient approaches for multi-objective sequential decision making. IJCNN 2014: 2323-2330.

[9] Rahmes, M., Pemble, R., Lemieux, G., Fox, K., "Multi-Dimensional Reward Volumes for Sensor Priority Strategies", IEEE CCNC 2014 Conference, January 10, 2014.

[10] Rahmes, M., Wilder, K., Yates, H., Fox, K., "Near Real Time Discovery and Conversion of Open Source Information to a Reward Matrix", WMSCI 2013, 12 July 2013.

[11] Rahmes, Delay, J., Cook, E., Hackett, J.; Optimal Multi Dimensional Fusion Model for Sensor Allocation and Accuracy Assessment; MILCOM 15, Oct 2015.

[12] Roy, O., Epistemic logic and the foundations of decision and game theory. **Journal of the Indian Council of Philosophical Research**, *27*(2), 2010, pp. 283–314.

[13] Wayne Winston, **Operations Research Applications and Algorithms** 4th. Edition, 2003.