12th International Conference on Information Fusion
Seattle, WA, USA, July 6-9, 2009

Track-Centric Metrics for Track Fusion Systems

Robert Canavan and Dr. Claire McCullough
Dept. of Computer Science and Engineering
University of Tennessee at Chattanooga
Chattanooga, TN, U.S.A.
claire-mccullough@utc.edu

William J. Farrell III
R&D Department
Adaptive Methods, Inc.
Centreville, VA, U.S.A.
jfarrell@adaptivemethods.com

Abstract - *In order to reduce the labor-intensive nature of evaluating quality of fused tracks in a multi-sensor, multi-platform environment, the Metrics Assessment System (MAS) was developed to provide heavily automated, real-time metric calculation and display. The MAS presents metrics in tables and graphs within the tactical display and provides overlays on the tactical display for focused assessment. Originally, MAS included standard kinematic metrics, but more recent work has generated additional metrics based on track-to-track versions of Track Purity, Target Effectiveness, and Assignment Accuracy, metrics derived from a confusion matrix analysis of the fused and known truth tracks, and Receiver Operating Characteristic Curves. Data mining tools were used to assess the effectiveness of candidate metrics for inclusion in the MAS. This paper describes the metrics, the evaluations, and the framework for metric assessment which this work has provided.*

Keywords: Tracking, track fusion, performance evaluation, data mining, metrics

1 Introduction

There are many situations in the Anti-submarine Warfare (ASW) domain where automated track fusion does not perform well, but the reasons for the sub-optimal performance are not completely understood. In order to improve performance, it is necessary for fusion system developers to understand situations where their algorithms fail and the cause of these failures. It can be a very labor-intensive process to perform this analysis after the fact on the full set of mission data; therefore, the Metrics Assessment System (MAS) was developed to support rapid, real-time testing and improvement assessment. The MAS capability provides graphical, tabular, and overlay presentations of defined metrics for a variety of Under Sea Warfare (USW) scenarios. As a result, data fusion developers can quickly perform quantitative assessment of their algorithms and gain insight into the reasons for poor performance as the system is processing data. This capability enables the developer to gain insight, which is often difficult to obtain during off-line analysis performed after the fact.

The MAS provides the framework for the display of metrics in real time. The metrics for data fusion analysis may be categorized into two groups. The first group is measures of performance (MOP), which quantify the capability of the system under test in performing the tasks for which it was designed. The second category of metrics is measures of effectiveness (MOE), which quantify the utility of this system as it relates to the successful completion of a mission or task [1]. The main focus of this work has been in selecting the most effective set of MOPs to include in the MAS.

The sections that follow first provide a brief description of the MAS, the metrics selected for possible inclusion in the system, and the process of evaluating metrics for possible inclusion using data mining tools, followed by conclusions and future work.

2 Overview of MAS

The MAS provides an abstract definition of a “metric” and “metric history” so that a variety of metrics can be implemented without any specific knowledge of the data fusion system being assessed. Once a developer has defined a metric and implemented it in the form of an algorithm, it can be made available to MAS by implementing the plug-in interface and adding the metric to an XML configuration file that is processed during MAS startup.

The MAS visualization component presents the operator with a list of metrics that are currently running and producing metrics values. An operator can select one of three charts and select a metric from the list to display that metric’s time series data. The time series data is updated in real-time as the metrics database is populated with new metric values. An operator can choose any metric and select the “Metric Details” tab for more information about the objects within the track picture. MAS also provides a “Detail History” tab, which shows the entire time-history for a selected metric and a selected fused track selected on the “Metric Details” tab. As a result, the three tabs shown on the MAS display (Metric List, Metric Details, and Detail History) allow the operator to drill down on track-centric metrics for a particular track to precisely determine the time at which a particular metric reaches a particular value.

The benefit of these time series representations is that an operator can easily correlate scenario events in real-time. Changes in metrics during target maneuvers, target crossing, or target fades can lend immediate, deep insight into fusion engine deficiencies.

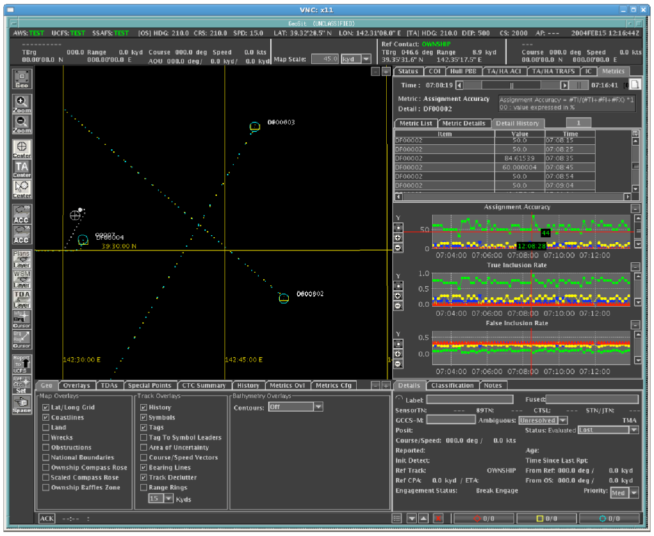

Figure 1. Example of MAS display with three metrics charts and metric tabs on the right.

In addition to the metric charting and details displays, MAS provides a geo-spatial overlay for the Area of Uncertainty (AOU) provided by a fusion system. Given the fused track covariance provided by a fusion system, MAS computes the 95% spatial AOU for a specified track and displays this AOU on the tactical geo-situational display in real-time as the fusion track is updated. In addition, the operator can select a "truth" track for comparison. With a "fusion" track and a "truth" track selected, MAS renders the AOU on the display and color-codes it depending upon whether the selected "truth" track is contained within the AOU of the selected "fusion" track [2].

3 Original Metrics

Several common kinematics-based track fusion metrics were already implemented within the MAS framework. In many cases, these metrics require knowledge of the track picture *ground truth*. This ground truth can be in the form of: (1) ground truth for a simulated data set, (2) ground truth reconstructed from a test event, or (3) ground truth provided by a trusted data source or sensor. These included:

- Total Number of "Fusion" Tracks
- Total Number of "Truth" Tracks
- Number of Source Tracks by Source
- Fusion Track Kinematic Components
- Fusion Track AOU Semi-Major Axis
- Completeness (Global Metric): % of "truth" tracks that lie within at least one "fusion" track AOU
- Spuriousness (Global Metric): % of "fusion" tracks that contain no "truth" tracks within their AOU
- Position Ambiguity (Global Metric): % of "fusion" tracks that have >1 "truth" track within its AOU
- Position Accuracy (Track Metric): Euclidian distance between the "best" fusion-truth track pair
- AOU Consistency (Track Metric): Position Accuracy divided by the AOU Semi-Major Axis

The latter two metrics require MAS to compute the "best" fusion-truth track pairing for all fusion tracks [3]. For more elaborate comparative metrics, this "best" fusion-truth track pair must be calculated routinely. Therefore, MAS periodically computes the "best" fusion-truth track pairings using the current "fusion" and "truth" track pictures for MAS collectors to use for the basis of their calculations. These pairings are stored in the metrics database for any metrics plug-in to access.

The fusion-truth track pairings are computed using the Modified Auction algorithm [4]. The Modified Auction algorithm is an optimal 2D assignment algorithm that applies the one-to-one assignment constraint such that a "fusion" track is assigned to no more than one "truth" track. For implementation of the Modified Auction algorithm, the Log-Likelihood Ratio cost function is used [5] and computed using the fusion track's AOU and residuals between fusion and truth tracks. From the results of the Modified Auction process, the Completeness and Spuriousness metrics can immediately be derived. Completeness is equal to the % of "truth" tracks that are assigned to a "fusion" track, while Spuriousness is equal to the % of "truth" track that are left unassigned to a "fusion" track. The "fusion" tracks that are assigned to "truth" tracks are subsequently used to compute the Position Accuracy and AOU Consistency metrics. These assignment pairings can be used to compute a variety of track-centric metrics, including those discussed in the remainder of this paper.

4 Candidate Metrics

4.1 Standard Metrics

In order to evaluate a set of potential metrics for inclusion in the MAS, a simulated scenario was run with programming code modified to calculate a confusion matrix and output the results to a text log file.

The terms used in other disciplines for confusion matrix analysis were modified for their usage in evaluation of data fusion for this ASW problem. The four status categories are True Inclusion (TI), False Inclusion (FI), True Exclusion (TX) and False Exclusion (FX), referring to whether a sensor track is, and should be, included in the fused track. These are the logical equivalents of true positive, false positive, true negative and false negative, respectively. Figure 2 is a chart of the confusion matrix with some of the terms used in other disciplines supplied for comparison to the terms used here.

		Signal, State of the World, **Truth Track**	
		Yes, True, Signal, **Included**	No, False, Noise, **Excluded**
Receiver, Response, **Fusion Track**	Yes, True, **Included**	Hit, Sensitivity, True Positive Fraction (TPF), **True Inclusion (TI)**	False Alarm, False Positive Fraction (FPF), **False Inclusion (FI)**
	No, False, **Excluded**	Miss, False Negative Fraction(FNF), **False Exclusion (FX)**	Correct Rejection, True Negative Fraction (TNF), **True Exclusion (TX)**

Figure 2. An example confusion matrix, including analogous terms from different disciplines. The terms used in this paper for data fusion evaluation are in bold [6][7][8].

The file generated by running the test scenario contained 1898 records that covered approximately two hours of scenario time after all of the sensors came on-line during system start up. This file was imported into a Microsoft Excel spreadsheet for off-line calculation and evaluation of candidate metrics. The intent of this process was to pre-screen candidate metrics and then add the most promising candidates to the MAS. The candidate metrics and their formulas, using the terminology from Figure 1, are:

Track Purity (TP) $= \frac{TI}{TI+FI+FX}$ (1)

Target Effectiveness (TE) $= \frac{\text{TI}}{\text{size of Fusion Track Set}}$ (2)

Assignment Accuracy (AA) $= \frac{TI}{\text{size of Truth Track Set}}$ (3)

Accuracy (ACC) $= \frac{TI+TX}{TI+TX+FI+FX}$ (4)

Specificity (SPC) $= \frac{TX}{TX+FI}$ (5)

Positive Predictive Value (PPV) $= \frac{TI}{TI+FI}$ (6)

Negative Predictive Value (NPV) $= \frac{TX}{TX+FX}$ (7)

False Discovery Rate (FDR) $= \frac{FI}{FI+TI}$ (8)

Matthews Correlation Coefficient (9)

$$(\text{MCC}) = \frac{(TI*TX)-(FI*FX)}{\sqrt{(TI+FI)*(TI+FX)*(TX+FI)*(TX+FX)}}$$

4.2 Receiver Operating Curve Metrics

The Receiver Operating Characteristic (ROC) curve graph classically consists of a plot of 1-Specificity vs. Sensitivity. These two values are synonymous with the False Inclusion Rate (FIR) and True Inclusion Rate (TIR) metrics [9].

True Inclusion Rate (TIR) $= \frac{TI}{TI+FX}$ (10)

False Inclusion rate (FIR) $= \frac{FI}{FI+TX}$ (11)

For purposes of calculating the area under the ROC curve (AUC), an approximation of the entire ROC curve is formed by using the three points (0,0), (FIR, TIR), and (1,1). This is necessary since the decision threshold is static for each point of analysis for the tracks and cannot be adjusted to generate the entire ROC curve. The area is then calculated with geometry, using the area of two triangles and a rectangle.

Area Under the ROC Curve (AUC) =

$$(\frac{1}{2}*(TIR*FIR))+(TIR*(1-FIR))+(\frac{1}{2}*(1-TIR)*(1-FIR))) \quad (12)$$

4.3 Goodness of Fit

An additional value, the Goodness of Fit (GOF), was imported from the MAS system's 2-D optimal assignment Auction Algorithm calculations for evaluation as a metric. This GOF is not a 'standard' formula, but functions as a scoring methodology for the Auction Algorithm.

Goodness of Fit (GOF) = LLR_Threshold -
Mahalanobis_Distance
–ln(determinant(track_covariance_prediction))
+ln(track_probability_of_detection) - ln(1
–track_probability_of_detection)

where LLR_Threshold = 20.5 (99.9% value for a 5-dimensional multi-variate Gaussian density) (13)

Mahalanobis distance is a measure used to quantify the differences between sets, where

Mahalanobis_Distance = (14)
transpose(residual_state_vector).inv
(track_covariance_prediction).residual_state_vector

with the residual_state_vector being the difference between the extrapolated track state and the "ground truth" track state.

4.4 Statistical Metrics

Additional statistical variations were calculated for some of the candidate metrics. The change in value (Δ) of TI, TX, FI, FX, TIR, FIR, ACC, SPC, PPV, NPV, FDR and MCC from one reading to the next were calculated and recorded. Also, a standard moving average of the last ten values and an exponential moving average, which gives more weight to recent information, of the last ten values were calculated for the Assignment Accuracy, Goodness of Fit (GOF) and Area Under the ROC curve (AUC) values.

Standard Moving Average of last 10 values

$$(\text{SMA}) = \frac{\text{Sum of last 10 values}}{10} \qquad (15)$$

Exponential Moving Average of last 10 values

$$(\text{EMA}) = ((\text{Current Value} - \text{SMA})*\text{r}) + \text{Previous EMA} \qquad (16)$$

where r is a weighting factor calculated from the sample size (n), as

$$r = \frac{2}{n+1} \qquad (17)$$

5 Data Mining as Metric Evaluator

In an effort to determine the usefulness of the candidate metrics, the metrics spreadsheet was fed into the data-mining tool SAS Enterprise Miner.

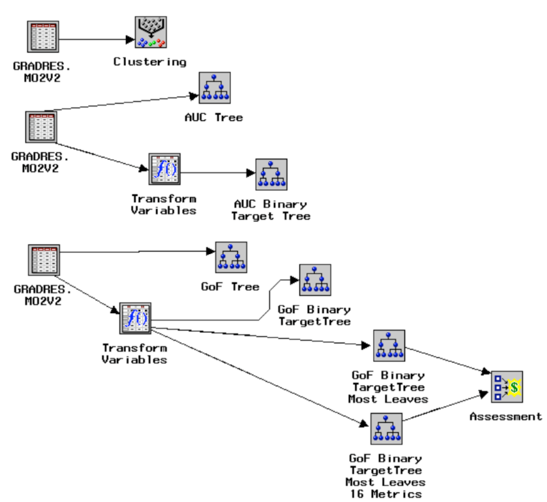

Figure 3. Chart of the flow of Data Mining Operations used in the analysis for this paper.

Enterprise Miner is one of a number of data-mining programs that can quickly perform calculations used to discover underlying patterns in large sets of data. Enterprise Miner allows the user to create data tables for processing. The data can then be explored to uncover the hidden relationships. Data can be modified to forms that are appropriate for different modeling techniques. A variety of modeling techniques is included to allow the user to find the method that will best predict the outcome. Finally, the assessment process allows the user to compare the different models' results to gain an understanding of their reliability [10]. It was hypothesized that this type of analysis can help to determine which candidate metrics contribute significant information and which do not.

In Figure 3, the GRADRES.MO2V2 node represents the metrics data for the 1898 records. The Transform Variables nodes partition the records into 'good' and 'bad' bins. The cluster and tree nodes perform that type of data mining analysis. The Assessment node performs comparisons of results.

5.1 Cluster Analysis

The first evaluation of the data was by clustering. The data points were plotted in 36 dimensions, (each candidate metric adding a dimension), and then analyzed to determine how many sub-groups emerged. In the clustering tool, sub-groups are areas of the plotted data space that have a high density of values. The distance of a value from the centroid of a cluster determines the cluster to which that value is assigned. By using an iterative process that moves the centroids and recalculates all of the distances from the data points to this revised centroid, the centroid locations which provide the best differentiation of the data can be determined. This analysis also provides information on the metrics which are most important in determining this differentiation. The simulation data cluster calculations resulted in two clusters that used Assignment Accuracy (AA) as the metric differentiating between them

5.2 Binary Tree Analysis

There are many different data mining models that could be used to discover other relationships within the data. A binary tree analysis was chosen since it produces an easily interpreted graph of the model's process and the relative importance of the different metrics used to build it. A binary tree divides the data into two groups based upon the metric that provides the most 'information gain' at that split. The information value for each split is calculated using the entropy (or information value) equation:

$$\text{entropy}\,(p_1,p_2) = -p_1*\log_2(p_1) - p_2*\log_2(p_2) \qquad (18)$$

with p being the estimated probability for each of the two sub-groups under test [9]. This type of tree analysis is an iterative process whereby the split with the highest information gain is found for each group. Then each of those sub-groups has the information gain recalculated, and the best split is performed for each of those, etc., until an ending condition is met. There are a number of choices for ending condition depending upon the situation and the amount of data. These restrictions can include things like the maximum depth of the chart, a minimum sub-group size required to perform a split, or a target statistical threshold.

The Area Under the ROC Curve (AUC) and Goodness of Fit (GoF) metrics were chosen as targets for different runs of the tree analysis. Since the data being utilized is from a simulation, a conservative approach in metrics evaluation was taken to avoid 'over fitting' the metrics to this one scenario. The binary tree analysis was performed twice for each of the target metrics. The first tree analysis predicts the value of the target metric. This uses the value of other metrics to derive the value of the target metric. The second tree analysis was for a binary 'good bin or bad bin' prediction. A cutoff value was chosen for each target variable such that records to one side of that value were considered 'good' while those on the other side of that value were considered 'bad.'

5.2.1 AUC Trees

The model for the value of the Area Under the ROC Curve was set up as follows. The individual counts for TI, TX, FI, and FX were excluded from consideration since AUC is derived from these values. In experimentation with these values included, their effect overrode the interaction of the other metrics and resulted in trees that split only using these values. The splitting criterion settings are:

F-test at significance level of 0.200
Minimum observations in a leaf – 5
Observations required for split search – 18
Maximum Number of branches from node – 2
Maximum depth of tree – 6

The model assessment measure was left on automatic, which allows the program to chose the best assessment method for the model based on its internal calculations, and the Sub-tree was set to most number of leaves (i.e. resulting in the most splits before the stopping criteria are met). The resulting tree is:

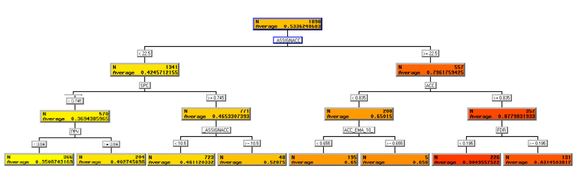

Figure 4. Tree Analysis AUC by Value Target

The good/bad bin analysis set the good bin as an AUC value equal to or above 0.80. All of the values below the 0.80 level were put into the 'bad' bin. This division is based upon guidelines for medical diagnostic tests for a 'good' or higher acceptability level of the test under evaluation [11]. The medical scale is used as a starting point for evaluation of data fusion since the potential downside risks of injury or death due to incorrect medical diagnosis appear similar to the potential downside risks of incorrect battlefield situational awareness.

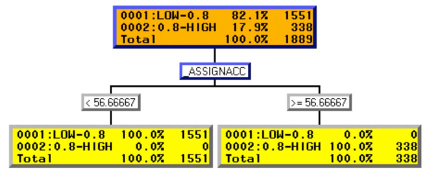

Figure 5. Tree Analysis AUC Good/Bad Bin

Note that, as in the cluster analysis, Assignment Accuracy values alone, labeled as _ASSIGNACC in Figure 4, can be used to differentiate the data into two sub-groups by dividing them into good and bad bins based on an Assignment Accuracy value above or below 56.66667. Since none of the values were misclassified, these two metrics can be considered to have a very strong correlation.

5.2.2 GoF Trees

The same type of tree analysis that was performed on the Area Under the ROC Curve (AUC) was repeated for the Goodness of Fit (GoF) readings to try to determine whether a relationship exists between GoF and the other candidate metrics. These runs were made with the same software settings as for the AUC tests, and the resulting binary tree for the value analysis is shown in Figure 6.

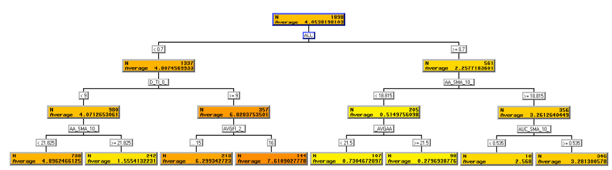

Figure 6. Tree Analysis Goodness of Fit– Value

For the good/bad bin analysis version of the tree analysis in Figure 7, the discrimination level between the bins was set at a GoF value of 1 (good being <=1 and bad being >1). This covers approximately the same percentage of readings, about 17%, as those above 0.80 in the Area Under the Curve Analysis. Since that portion of the data is considered "good or better" in the AUC analysis, a similar fraction of the data is used in this analysis. No attempt was made to determine the amount of overlap within these datasets.

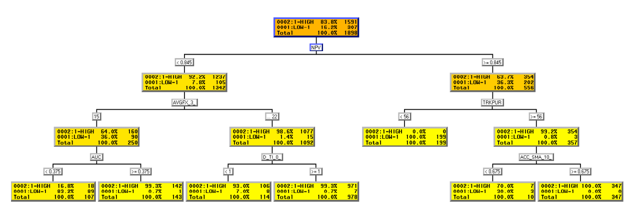

Figure 7 - Tree Analysis Goodness of Fit– Good/Bad Bin

Since the intent of binary tree modeling is to provide the largest gain of information at each split, the metrics that are included in the trees provide the most discrimination.

6 Results of Data Mining Analysis

From the list of all 36-candidate metrics, the following were included in the trees, with the number of inclusions noted beside each. Related categories are listed together:

Metric	Inclusions
Assignment Accuracy (AA)	3
SMA of Assignment Accuracy (last 10 values)	2
Average of Assignment Accuracy	1
Accuracy (ACC)	2
SMA of Accuracy (last 10 values)	1
EMA of Accuracy (last 10 values)	1
Delta of True Inclusion (d-TI)	2
Average False Inclusion (FI) in percentage	1
True Exclusion (TX) Count	1
Specificity (SPC)	1
Positive Predictive Value (PPV)	1
Negative Predictive Value (NPV)	1
False Discovery Rate	1
Area Under the ROC Curve (AUC)	1
SMA of AUC (last 10values)	1
Track Purity (TP)	1

7 Metrics Chosen for MAS

7.1 Track-Centric Metrics

The metrics included in section 6 appear to be the ones that have the potential to provide the most useful information in the fusion assessment process, but with a desire for concise and consistent code, decisions were made as to what to include in MAS. Since standard moving average was used more than exponential moving average in the trees, only the standard moving average was included. Since only one of the deltas was used, the whole category was removed as metric candidates. The previously included metrics (AA, TP, TE) were included for consistency with the earlier fusion algorithm analysis by Dr. Chidambar Ganesh [12] on USW problems. Goodness of Fit was included, since it appears to have the potential to reveal variation that the confusion matrix derived metrics might not reveal. Finally, because the Area Under the ROC Curve is calculated from the True Inclusion Rate and False Inclusion Rate, they were also included. The final list of metrics included in the MAS as an initial track evaluation framework is:

Assignment Accuracy
Moving Average of Assignment Accuracy
Standard Moving Average of Assignment Accuracy (last 10 measures)
Target Effectiveness
Moving Average of Target Effectiveness
Track Purity
Moving Average of Track Purity
Confusion matrix in percentages (TI, FI, TX, FX)
True Inclusion Rate
False Inclusion Rate
Area Under the ROC Curve
Standard Moving Average of Area Under the ROC Curve (last 10 values)
Accuracy
False Discovery Rate
Goodness of Fit
Standard Moving Average of Goodness of Fit (last 10 values)

In an effort to understand the amount of information that might be lost by reducing the set of metrics from 36 to 16, the binary tree analysis for Goodness of Fit - Good/Bad Bin was recalculated using only these 16 metrics. The ROC curve generated by the reduced metrics set was then compared to the original 36 metric ROC curve in order to see how much the area under these curves changed due to the removal of 20 metrics. The difference in area under the curve gives an indication of how much the quality of the metrics model has changed.

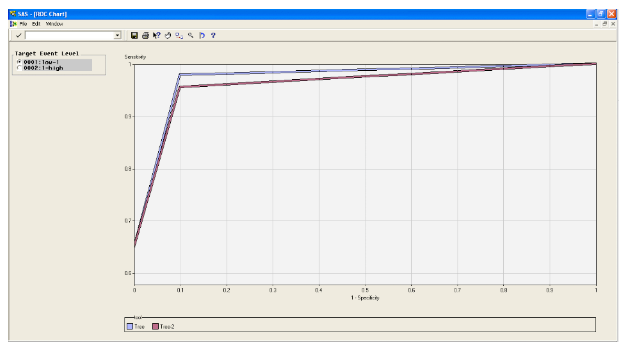

Figure 8 - ROC Curve Chart for All 36 Candidate Metrics (Tree- Blue) and 16 Candidate Metrics (Tree-2 – Red)

Figure 8 demonstrates the information loss due to the reduction in the number of metrics. This loss is from a reduction in the Sensitivity. In other words, there may be some 'hits' that are not included. There was no change in the 1-Specificity. Using geometry to estimate the area under the curves yields approximately 0.94 for the 36-metric curve and 0.93 for the 16-metric curve. This represents roughly a 1% drop in area due to the reduced set of metrics, yet both are still above the 'excellent' level of 0.90 in the medical diagnostic test evaluation scale [11]. Therefore, the amount of information loss seems to be acceptable in relation to the amount of complexity removed. This complexity includes computer code writing time, increasing the length of the metrics list on the MAS simulation display, and the learning curve for understanding all of the metrics by fusion system evaluators. Future work with these metrics in the evaluation process may reveal that some of these metrics prove to be more valuable than others in the analysis of real world tracks, and some of the excluded metrics may be revisited.

7.2 Sensor-Centric Metrics

It is also desirable to calculate metrics for individual sensors so that their usefulness in contributing to the correct fused track could be measured. From the work on track analysis, it seems clear that Assignment Accuracy is a metric that could provide a significant amount of information as to the usefulness of a source. The calculation of the confusion matrix quickly demonstrates what types of errors the inclusion or exclusion of the sensor tracks is producing. The Area Under the ROC Curve gives a single metric to describe to overall effect of these errors. TIR and FIR are used in the calculation of AUC, so they need to be included as well. So for each sensor, the following metrics are calculated and updated on a moving average basis:

Confusion Matrix –TI, FI, TX, FX
True Inclusion Rate (TIR)
False Inclusion Rate (FIR)
Area Under the ROC Curve (AUC)
Assignment Accuracy (AA)

The changes were made to the MAS to include the noted metrics for the track and sensor level. The metrics were made available to the display as well as being recorded in the metrics log text file.

8 Results

8.1 Track Level Review

While running the MAS system, the newly included metrics are shown in the display graphs and lists under the metric tab.

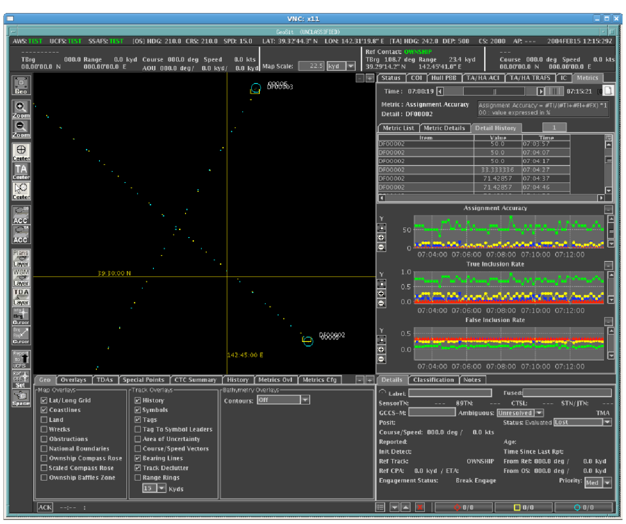

Figure 9. Screen Shot of Test Scenario showing low Assignment Accuracy point for Fusion Track DF00002.

At this stage in the metrics development process, the output to the metrics log is intended for use as a reference for investigation of events noted in the display. For example, in the screen shot in Figure 9, there is a point in the graph of Assignment Accuracy (AA) for fusion track #DF00002 (the green line on the metrics graph) that is lower than the variation seen in the rest of the graph. This target is a surface contact represented by the blue circle in the lower right quadrant of the map display. The truth track associated with this fusion track is the yellow open rectangle in the same area. The detail history shows that this point occurred at simulation time 07:04:27 with an AA value of 33.333336. To make an assessment of the low Assignment Accuracy point, the information can be quickly found in the metric log. Information from other points on that same track around the same time can also be extracted for comparison, as shown in Figure 10.

Previous Point	Low Point	Next Point
Current FI set...	Current FI set...	Current FI set...
[SID/STN] = [37/0]	**[SID/STN] = [37/0]**	**[SID/STN] = [37/0]**
[SID/STN] = [21/3]	[SID/STN] = [21/3]	**[SID/STN] = [24/3]**
[SID/STN] = [4/2]	[SID/STN] = [4/2]	
[SID/STN] = [24/3]	[SID/STN] = [3/2]	
	[SID/STN] = [2/2]	
	[SID/STN] = [24/3]	
Current FX set...	Current FX set...	Current FX set...
[SID/STN] = [24/1]	**[SID/STN] = [24/1]**	**[SID/STN] = [24/1]**
[SID/STN] = [4/1]	[SID/STN] = [4/1]	**[SID/STN] = [40/1]**
[SID/STN] = [40/1]	[SID/STN] = [3/1]	
[SID/STN] = [21/1]	[SID/STN] = [2/1]	
	[SID/STN] = [40/1]	
	[SID/STN] = [21/1]	

Figure 10. Summary of metric data from the metrics log for the low Assignment Accuracy point on fusion track DF00002.

Note that in Figure 10, the sensor tracks that are common to all of the points are in bold. The SID is the sensor identification number and the STN is the track number for that sensor.

A discovery made during the test scenario evaluation was the inclusion of sensor tracks from a sensor labeled #37, which is not in the list of fusible sources for this simulated scenario. It also includes, as is shown in the first line of Figure 10, an invalid sensor track number of 0. All sensor tracks for this simulation should have track numbers of one or above. The discovery of an anomaly in the MAS simulation scenario which had not been identified in multiple earlier uses of the same scenario demonstrates the potential effectiveness of the metrics code in discovering irregularities in the MAS system or in the way in which tracks are fused by the fusion engine.

8.2 Sensor Level Review

After running the MAS system test simulation for approximately half an hour of scenario time, the metrics data from each sensor was reviewed. The quality of the information provided by the sensors seemed to fall into three categories based on the Area Under the ROC curves (AUC). They are either above 0.50, about 0.50 or below 0.50.

Sensor #	23	10	4	22	30	24	40	2	3	41	29	21
Sensor Type	HA ACTIVE	Radar	PAFS PREC NB TRK	HA ACI	TRAFS TA DCL TAP	TA ACTIVE	HPPFS PB5 CF	Pafs Prec TRK	TA PBB ACF	HPPFS PBB PT	TRAFS HA DCL TAP	TA ACI
Metrics												
TI	15.94	8.81	8.85	8.66	6.49	6.61	**0.00**	5.01	5.05	4.67	4.11	3.37
FI	9.06	16.19	15.83	16.30	18.03	18.39	0.00	19.67	19.79	20.33	20.89	21.47
TX	65.94	59.55	59.17	58.70	56.97	56.61	75.00	55.33	55.21	54.67	54.11	53.53
FX	9.06	15.45	16.15	16.34	18.51	18.39	25.00	19.99	19.95	20.33	20.89	21.63
TIR	0.64	0.36	0.35	0.35	0.26	0.26	0.00	0.20	0.20	0.19	0.16	0.13
FIR	0.12	0.21	0.21	0.22	0.24	0.25	0.00	0.26	0.26	0.27	0.28	0.29
AUC	**0.76**	**0.57**	**0.57**	**0.56**	**0.51**	**0.51**	**0.50**	**0.47**	**0.47**	**0.46**	**0.44**	**0.42**
AA	46.81	21.77	21.69	20.98	15.08	15.24	**0.00**	11.21	11.27	10.30	8.97	7.24

Figure 11. Metrics on Sensors from test scenario.

The sensors with ROC areas over 0.50 are included correctly at a higher rate than they are included incorrectly. For the sensors that are around the 0.50 range, the fusion engine includes the sensor correctly and incorrectly at about the same rate, the same level of performance that a random guess would generate. The sensors with AUC results that are under 0.50 are falsely included at a higher rate then they are truly included. Therefore, if the decision to include or not include was calculated and then inverted, that sensor would be included correctly a greater percentage of the time. This analysis does not make any judgments on the value of the sensor's data, just the fusion engine's decision to include that data in the fusion track or not. In future work, this could lead to the development of intelligent fusion techniques that could preferentially include or exclude sensor tracks based upon each sensor's historical record of accuracy. For example, while sensor #40 has an AUC of 0.50, it also has Assignment Accuracy of 0, meaning that the fusion engine never includes it properly. If this type of observation is repeated in multiple scenarios either for specific sensors or categories of similar sensors, then these common threads can be flagged as specific areas for improvements.

9 Concluding Remarks

While the addition of track-centric metrics to the MAS system has already proved beneficial in observing trends in the way in which tracks from specific sensors are included by the fusion engine, and in discovering glitches in the simulation scenario, much work remains to be done. The next step in the process is to run the MAS with a wide variety of simulated scenarios to gain a more complete set of data on which the performance of the current, or other, fusion engines can be compared, since the data can now be evaluated through the use of metrics on a per-track and per-sensor basis to look for interesting trends during the run. After thorough evaluation on simulated data, the validated metrics can be used with sea test data to give an evaluation of available fusion techniques more representative of actual ASW operational performance.

Additionally, other metrics should be investigated to supplement or improve the metrics already included, such as inclusion of new MOE metrics to supplement the existing MOPs.

10 References

[1] Hall, David L. and Sonya A.H. McMullen, *Mathematical Techniques in Multisensor Data Fusion* 2nd Edition, Artech House, Boston, 2004.

[2] Farrell, William J. III, Claire McCullough, and Robert Canavan, "Automated metrics assessment system for track fusion," SPIE Defense, Security, and Sensing. Orlando, Florida, *Conference 7336 - Proceedings of SPIE* Volume 7336, April 2009.

[3] Single Integrated Air Picture (SIAP) System Engineering Task Force, "Single Integrated Air Picture (SIAP) Attributes," Technical Report 2003-029, Version 2.0, 2003.

[4] Blair, W.D. and Y. Bar-Shalom, *Multitarget-Multisensor Tracking Applications and Advances,* Volume III, Artech House, Boston, 2000.

[5] Blackman, Samuel and Robert Popoli, *Design and Analysis of Modern Tracking Systems*, Artech House, Boston, 1999.

[6] Wickens, Christopher D. and Justin G. Hollands, *Engineering Psychology and Human Performance,* 3rd Edition, New Jersey, Prentice Hall, 2000.

[7] Heeger, David, "Signal Detection Theory Handout," 1998, Department of Psychology, Stanford University, http://www-psych.stanford.edu/~lera/psych115s/notes/signal/, last accessed 1/11/09.

[8] Hopley, Lara and Jo van Schalkwyk, "The Magnificent ROC," 3/25/07, Anaesthetist.com, http://www.anaesthetist.com/mnm/stats/roc/Findex.htm, last accessed 1/9/09.

[9] Witten, Ian H. and Eibe Frank, *Data Mining, Practical Machine Learning Tools and Techniques,* 2nd Edition, Morgan Kaufmann, New York, 2005.

[10]SAS Institute Inc., *Getting Started with SAS® Enterprise Miner™ 5.3,* SAS Institute Inc., Cary, NC, 2008. http://support.sas.com/documentation/onlinedoc/miner/getstarted53.pdf, last accessed 3/13/09.

[11]Tape MD, Thomas G., "The Area Under an ROC Curve," University of Nebraska Medical Center, http://gim.unmc.edu/dxtests/roc3.htm, last accessed 1/13/09.

[12]Ganesh, Chidambar, "Undersea Warfare Decision Support System, Fusion Assessment: DSS Technical Review [PowerPoint Slides]," Chattanooga, TN, 9/17/08.