Information theoretic approach for performance evaluation of multi-class assignment systems

Citation: Ryan S. Holt, Peter A. Mastromarino, Edward K. Kao, and Michael B. Hurley (2010). Information theoretic approach for performance evaluation of multi-class assignment systems. Proc. SPIE 7697: 76970R/1-12. ©2010 COPYRIGHT SPIE--The International Society for Optical Engineering

As Published: http://dx.doi.org/10.1117/12.851019

Publisher: SPIE

Persistent URL: http://hdl.handle.net/1721.1/58584

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Information theoretic approach for performance evaluation of multi-class assignment systems

Ryan S. Holt, Peter A. Mastromarino, Edward K. Kao, Michael B. Hurley[*†]
MIT Lincoln Laboratory, 244 Wood Street, Lexington, MA, USA 02420-9185

ABSTRACT

Multi-class assignment is often used to aid in the exploitation of data in the Intelligence, Surveillance, and Reconnaissance (ISR) community. For example, tracking systems collect detections into tracks and recognition systems classify objects into various categories. The reliability of these systems is highly contingent upon the correctness of the assignments. Conventional methods and metrics for evaluating assignment correctness only convey partial information about the system performance and are usually tied to the specific type of system being evaluated. Recently, information theory has been successfully applied to the tracking problem in order to develop an overall performance evaluation metric. In this paper, the information-theoretic framework is extended to measure the overall performance of any multi-class assignment system, specifically, any system that can be described using a confusion matrix. The performance is evaluated based upon the amount of truth information captured and the amount of false information reported by the system. The information content is quantified through conditional entropy and mutual information computations using numerical estimates of the association probabilities. The end result is analogous to the Receiver Operating Characteristic (ROC) curve used in signal detection theory. This paper compares these information quality metrics to existing metrics and demonstrates how to apply these metrics to evaluate the performance of a recognition system.

Keywords: Information Theory, measures of performance, recognition systems, classification

1. INTRODUCTION

With the increasing number of sensors being employed by the Intelligence, Surveillance, and Reconnaissance (ISR) community, the amount of sensor data that analysts are required to exploit continues to grow. Automated exploitation systems have been developed in order to aid the analyst by processing large amounts of data and providing alerts or cues to particularly significant or otherwise unusual data subsets. Such systems most often solve multi-class assignment problems, wherein the system attempts to assign objects to one of several (possibly many) truth classes. Tracking and recognition systems are both examples of systems that attempt to solve the multi-class assignment problem. A rigorous and comprehensive method for evaluating the performance of such systems is essential if they are to have any benefit for the end user.

In general, the end users typically have no difficulty articulating what they want in a tracking or recognition system and often can provide narratives of what they expect it to do. These descriptions suggest the need for a utility scoring function; unfortunately, the users seldom have sufficient knowledge of all the costs and benefits of the system to allow for such a utility analysis to be conducted. The system behaviors that are articulated often relate to different pathologies within the problem space, such as missed or false tracks in tracking systems and misclassifications in recognition systems. In order to capture these various behaviors, a number of performance metrics have been proposed by others through the years [1-6]. However, each metric only addresses a subset of the behaviors, and a single behavior frequently affects multiple metrics. To obtain an overall system performance score, these metrics have often been combined through an ad-hoc weighted sum with the weights being adjusted to reflect the users' unique needs. However, choosing a set of weights that is fair for algorithm comparison and that simultaneously reflects the requirements of a given application can be extremely difficult, especially when the metrics are in different units and are often highly correlated.

[*] {rholt, peterm, edward.kao, hurley}@ll.mit.edu

[†] This work was sponsored by the Department of the Air Force under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Air Force.

Signal Processing, Sensor Fusion, and Target Recognition XIX, edited by Ivan Kadar,
Proc. of SPIE Vol. 7697, 76970R · © 2010 SPIE · CCC code: 0277-786X/10/$18 · doi: 10.1117/12.851019

An overall scoring function based upon a theoretical foundation eliminates the problems of selecting ad-hoc weights. Recently, an information theoretic approach has been taken in the evaluation of both binary classifiers [7] and tracking systems [8]. The metrics presented in [8], truth information completeness and false information ratio, were particularly successful in providing a comprehensive evaluation of tracking performance. They were effective in capturing the effects of common track pathologies, allowing for tracker comparisons and parameter tuning in order to optimize tracker performance.

Given that tracking is a subset of the multi-class assignment problem, this paper generalizes the pair of metrics presented in [8] and demonstrates their application to any multi-class assignment system. We begin, in Section 2, by reviewing some evaluation metrics currently used and discuss their limitations. In Section 3, we describe the information quality metrics for the general multi-class assignment problem and compare them to the metrics presented in Section 2. Next, in Section 4, we show two examples how these metrics are used to evaluate recognition systems, parameter-tune classifiers, and describe the uncertainty on the performance evaluation estimate. Finally, in Section 5, we conclude our remarks and describe areas of future research.

2. EXISTING EVALUATION METRICS

In this section we review some of the more common existing methods for evaluating the performance of multi-class assignment systems. In general, the evaluation process involves running a certain number of test instances, each drawn from one of *N* possible truth classes, through the system and having it assign them each to one of *M* system classes, resulting in what is often referred to as an association or confusion matrix. Note that any $N \times M$ matrix can be made square by simply appending empty rows or columns; therefore, for convenience in the mathematical expressions, we assume $M = N$ and the confusion matrix to be square. Let $D(t,s)$ denote the confusion matrix output of a given system, where *t* and *s* denote one of *N* possible truth and system class labels, respectively. Let $|D|$ denote the cardinality of *D*, i.e., the total number of instances used to compute the confusion matrix. For purposes of "scoring" confusion matrices, it is convenient to express them in terms of joint probability distributions by simply dividing by the total number of instances, i.e., $P_{t,s} = D(t,s)/|D|$. The probability of a given truth class *i*, $P_t(i)$, or system class prediction *j*, $P_s(j)$, is then obtained by marginalizing over all system classes and truth classes, respectively:

$$\begin{aligned} P_t(i) &= \sum_{k=1}^{N} P_{t,s}(i,k) \\ P_s(j) &= \sum_{k=1}^{N} P_{t,s}(k,j). \end{aligned} \tag{1}$$

By far the simplest and most intuitive scoring method is to compute the *accuracy* metric by summing along the diagonal of the joint probability distribution:

$$\text{accuracy} = \sum_{k=1}^{N} D(k,k)/|D| = \sum_{k=1}^{N} P_{t,s}(k,k). \tag{2}$$

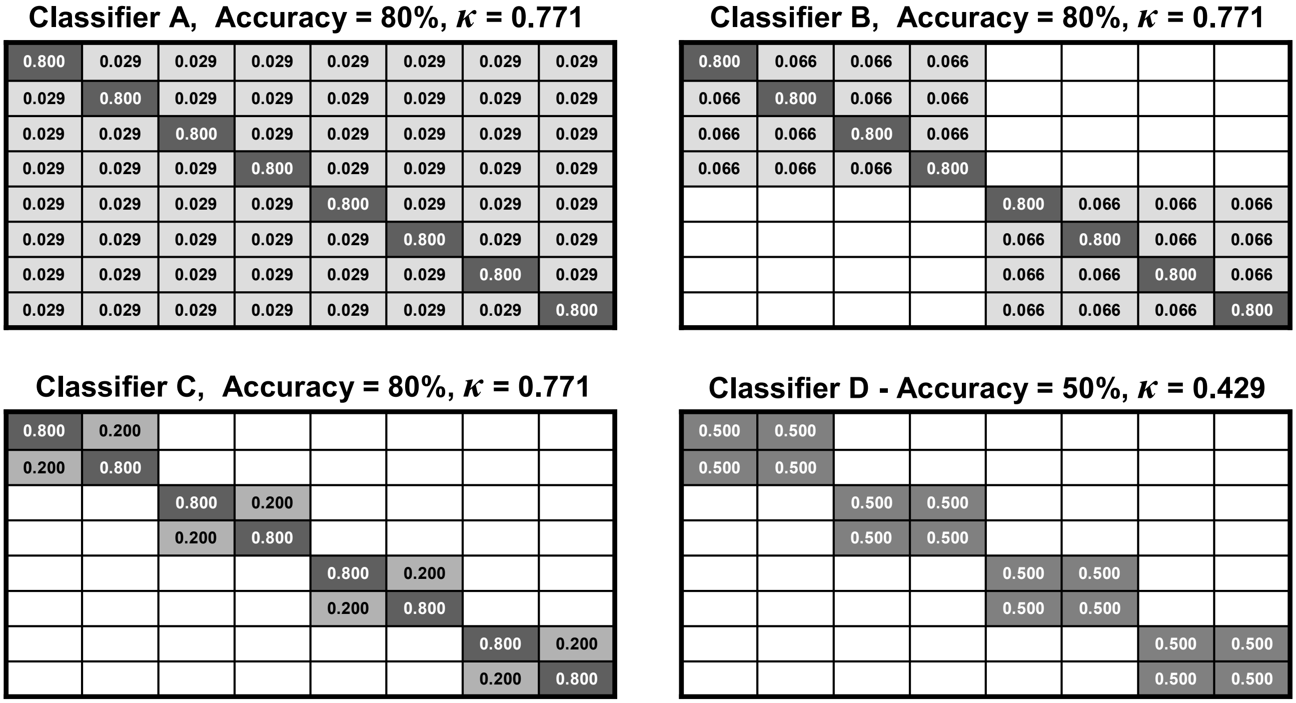
Accuracy is simply the fraction of correct assignments over all instances. The closely related *kappa statistic κ* modifies the accuracy metric by taking into account the number of correct classifications expected to occur purely by chance [1]. In this case, the truth class *t* and the system class prediction *s* are independent random variables, and no mutual information exists between them. The expectation value of the random-chance joint probability distribution is:

$$\hat{P}_{t,s}(i,j) = P_t(i) \cdot P_s(j). \tag{3}$$

The kappa statistic κ can then be written as a function of the accuracy metrics for the observed and random-chance joint probability distributions:

$$\kappa = \frac{\sum_{k=1}^{N} P_{t,s}(k,k) - \sum_{k=1}^{N} \hat{P}_{t,s}(k,k)}{1 - \sum_{k=1}^{N} \hat{P}_{t,s}(k,k)} . \qquad (4)$$

Both the accuracy and kappa statistics are metrics that are maximized in order to optimize system performance. However, neither alone provides a comprehensive performance evaluation, as can be seen by considering the confusion matrices in Figure 1. These confusion matrices represent the output of four notional classifiers A, B, C and D. Classifiers A – C all have identical accuracy and kappa statistics, even though there are clear differences in their performance. And while classifier D has a poor accuracy and kappa rating compared to the others, it is always able to narrow down the classification to at most two classes. By contrast, when A is wrong, it is equally likely to be wrong in any class.



Classifier A, Accuracy = 80%, κ = 0.771

0.800	0.029	0.029	0.029	0.029	0.029	0.029	0.029
0.029	0.800	0.029	0.029	0.029	0.029	0.029	0.029
0.029	0.029	0.800	0.029	0.029	0.029	0.029	0.029
0.029	0.029	0.029	0.800	0.029	0.029	0.029	0.029
0.029	0.029	0.029	0.029	0.800	0.029	0.029	0.029
0.029	0.029	0.029	0.029	0.029	0.800	0.029	0.029
0.029	0.029	0.029	0.029	0.029	0.029	0.800	0.029
0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.800

Classifier B, Accuracy = 80%, κ = 0.771

0.800	0.066	0.066	0.066				
0.066	0.800	0.066	0.066				
0.066	0.066	0.800	0.066				
0.066	0.066	0.066	0.800				
				0.800	0.066	0.066	0.066
				0.066	0.800	0.066	0.066
				0.066	0.066	0.800	0.066
				0.066	0.066	0.066	0.800

Classifier C, Accuracy = 80%, κ = 0.771

0.800	0.200						
0.200	0.800						
		0.800	0.200				
		0.200	0.800				
				0.800	0.200		
				0.200	0.800		
						0.800	0.200
						0.200	0.800

Classifier D - Accuracy = 50%, κ = 0.429

0.500	0.500						
0.500	0.500						
		0.500	0.500				
		0.500	0.500				
				0.500	0.500		
				0.500	0.500		
						0.500	0.500
						0.500	0.500

Figure 1. Illustrative example of four different classifiers and their confusion matrices. The percent accuracy and kappa statistic fail to capture all the structural information contained within the confusion matrix.

Additional metrics have been introduced that, among other things, attempt to provide an indication of the uncertainty connected to a given classifier's predictive ability, i.e., $P_{t|s}$. These metrics descend from detection theory and are thus generally defined in terms of *true / false* and *positives / negatives* [2]. Most readily applied to binary classification problems, they can easily be generalized to the multi-class assignment problem by considering it as a series of binary classification problems where the overall metrics are a weighted average of the individual class metrics. The following is a list of the most common metrics adapted for the multi-class assignment problem:

$$\text{True Positive Rate (TPR)} = \sum_{k=1}^{N} P_{t,s}(k,k) , \qquad (5)$$

$$\text{False Positive Rate (FPR)} = \sum_{k=1}^{N} P_t(k) \left(\frac{P_s(k) - P_{t,s}(k,k)}{1 - P_t(k)} \right) , \qquad (6)$$

$$\text{Positive Predictive Value (PPV)} = \sum_{k=1}^{N} P_t(k)\left(\frac{P_{t,s}(k,k)}{P_s(k)}\right), \tag{7}$$

$$\text{Negative Predictive Value (NPV)} = \sum_{k=1}^{N} P_t(k)\left(\frac{1-P_s(k)-P_t(k)+P_{t,s}(k,k)}{1-P_s(k)}\right), \tag{8}$$

$$\text{Rand index} = \sum_{k=1}^{N} P_t(k)\big(1-P_s(k)-P_t(k)+2P_{t,s}(k,k)\big). \tag{9}$$

Note that each of the above metrics has many aliases. TPR is often referred to as recall or sensitivity, FPR as false alarm rate or fall-out, PPV as precision, and Rand index [3] as accuracy in binary classification. Other metrics can be formed by considering different algebraic permutations of these fundamental expressions. However, none of them are able to adequately discriminate between classifiers A – D of Figure 1. For classifiers A – C the TPR, FPR, PPV, NPV, and Rand index values are 0.8, 0.029, 0.8, 0.97, and 0.95, respectively. Since each classifier shows identical values, no metric that is a combination of these rates can possibly distinguish between them. The same metrics for classifier D are 0.5, 0.071, 0.5, 0.93, and 0.88, respectively.

A metric sometimes used to evaluate overall system performance for binary classifiers is known as the *F-score* [4], which incorporates the TPR and the PPV. For multi-class systems, the weighted average of the F-scores is computed as:

$$\text{F-score} = \sum_{k=1}^{N} P_t(k)\frac{2P_{t,s}(k,k)}{P_t(k)+P_s(k)}. \tag{10}$$

The F-score for classifier D is 0.5, while classifiers A – C all have F-scores of 0.8. By this measure, then, classifier D is again ranked as the worst performer of all, and we are left unable to discriminate between classifiers A – C. The fundamental problem stems from the fact that we have turned a multi-class assignment problem into a series of binary classification problems, ignoring the subtle differences that may exist between specific types of misclassifications.

Loss functions attempt to address this deficiency by explicitly enumerating the costs of all relevant truth-to-system association hypotheses $t \mapsto s$ in the form of a cost matrix $C_{t,s}$ [5]. Given a loss function (i.e., cost matrix) and a joint probability distribution, system performance can be then defined as the expectation value of the costs:

$$\ell = \sum_{i,j=1}^{N} P_{t,s}(i,j)\cdot C_{t,s}(i,j). \tag{11}$$

The task of choosing an appropriate loss function is application-specific.* However, several standard loss functions exist that are appropriate to consider for multi-class assignment problems. The first is the *quadratic loss function* [5]:

$$C_{t,s}(i,j)_{quad} = \sum_{k=1}^{N} \big(P_{t,s}(i,k)-\delta(i,k)\big)^2. \tag{12}$$

The *linear* (or *absolute error*) *loss function* can also be considered when it is felt that the quadratic loss function penalizes misclassifications too severely [5]:

$$C_{t,s}(i,j)_{lin} = \sum_{k=1}^{N} \left|P_{t,s}(i,k)-\delta(i,k)\right|. \tag{13}$$

Both the quadratic and linear loss functions measure a mix of accuracy and precision, trying to minimize the difference between the observed joint probability distribution and the identity matrix $\delta(i,j)$ (i.e., $\delta(i,j)=1$ if $i=j$, 0 otherwise).

* In general, the only constraint a loss function must satisfy is that $C_{t,s}(k,k) < C_{t,s}(k, s \neq k)\,\forall k$; otherwise, correct classifications will be more (or, at best, equally) costly as misclassifications, and the "classifier" will hardly be deserving of the name.

The *informational loss function*, like the accuracy and kappa statistics, provides a measure only of a classifier's accuracy [6]. Here the cost is equal to the negative log-likelihood of a correct classification for each truth class:

$$C_{t,s}(i,j)_{inf} = -\log P_{t,s}(i,i). \tag{14}$$

Finally, the so-called *"0-1" loss function* recovers the accuracy performance metric of Equation (2), modulo a sign [5]. Here the cost of a correct classification is –1 while that of a misclassification is 0, the negative sign being necessary to preserve the correspondence between lower costs and better system performance.

$$C_{t,s}(i,j)_{0\text{-}1} = -\delta(i,j). \tag{15}$$

Table 1 lists the performance metrics for each of the classifier confusion matrices in Figure 1 using the four loss functions described above. The linear, informational, and "0-1" loss functions are unable to distinguish between classifiers A – C, even though there are obvious differences in the classifiers' results. The quadratic loss function is able to distinguish between classifiers A, B, and C, and ranks them 1, 2, 3, respectively, from best to worse. This is not surprising, since its goal is to minimize the difference between the joint probability distribution and the identity matrix, yet highly suspect, since it would seem that classifier C does the best job of both correctly classifying instances when it can and eliminating possibilities when it cannot. It's interesting to note that all four loss functions rank classifier D as having the worst performance, even though it is able to discriminate between more of the hypothesis space than classifier A, for instance.

Clearly none of the metrics described above is satisfactory, mainly owing to the fact that none of them relate directly to the specific goal of the multi-class assignment problem. A metric grounded in information theory, specifically tailored to the multi-class assignment problem, and able to measure both system accuracy and precision in a more holistic framework is sorely needed. The authors of [7] introduce one such metric, which they term the *normalized error to information ratio*, denoted here by ξ:

$$\xi = \alpha \cdot \frac{1 - \sum_{k=1}^{N} P_{t,s}(k,k)}{\sum_{j=1}^{N} P_s(j) \sum_{i=1}^{N} P_{t|s}(i\mid j) \log P_{t|s}(i\mid j) - \sum_{k=1}^{N} P_t(k) \log P_t(k)}. \tag{16}$$

The numerator is the classification error rate, while the denominator is the mutual information or the entropy added by the classifier system, reflecting the uncertainty of a truth class being correct given a specific system declaration [7]. The constant α is a tunable parameter. This metric is sensitive to both accuracy and precision. Setting $\alpha = 1$, this metric evaluates the performance of classifiers A, B, C, and D of Figure 1 to be 0.168, 0.147, 0.127, and 0.361, respectively. Thus the ranking is C, B, A, D, from best to worst, and again we find that classifiers A – C are all preferable to D. The motivation behind considering the ratio in Equation (15) is uncertain and the underlying question of whether this represents the best way of evaluating a classifier remains unanswered. An alternative information theoretic approach for performance evaluation of multi-class assignment systems is introduced in the next section.

Table 1. Performance evaluations for the classifier confusion matrices shown in Figure 1 for four commonly used loss functions: linear, quadratic, information, and "0-1".

Loss function	**A**	**B**	**C**	**D**
Linear	0.925	0.925	0.925	1.00
Quadratic	0.8101	0.8102	0.8106	0.8828
Informational	2.30	2.30	2.30	2.77
"0-1"	-0.80	-0.80	-0.80	-0.50

3. INFORMATION THEORETIC METRICS

Information theory is a useful framework for quantifying the amount of information in a given system and therefore can be used to measure the overall performance. In the paper by Kao et al. [8], information theory was applied to the evaluation of tracking systems. The main contributions of the paper were two metrics, truth information completeness $f(T;S)$ and false information ratio $r(S\mid T)$, which were successful in describing the overall performance of different tracking systems in addition to capturing all the different track pathologies. Since tracking systems are a specialized subset of multi-class assignment systems, in this section, we extend the information metrics to the general case. In addition, we show the comparison between these metrics and the metrics listed in the previous section.

In the tracking problem, system output tracks are associated together with truth tracks resulting in a matrix form. When normalized, this association matrix describes the joint probability function of the truth tracks with the system tracks. The joint probability function is then used to derive the five information measures shown in Figure 2 [9]: truth entropy $H(T)$ representing the total amount of truth information, system track entropy $H(S)$ representing the total amount of system information, mutual information $I(T;S)$ representing the amount of matching information between truth classes and system classes, truth conditional entropy $H(T\mid S)$ representing the amount of truth information missed by the system, and system conditional entropy $H(T)$ representing the amount of false information introduced by the system.

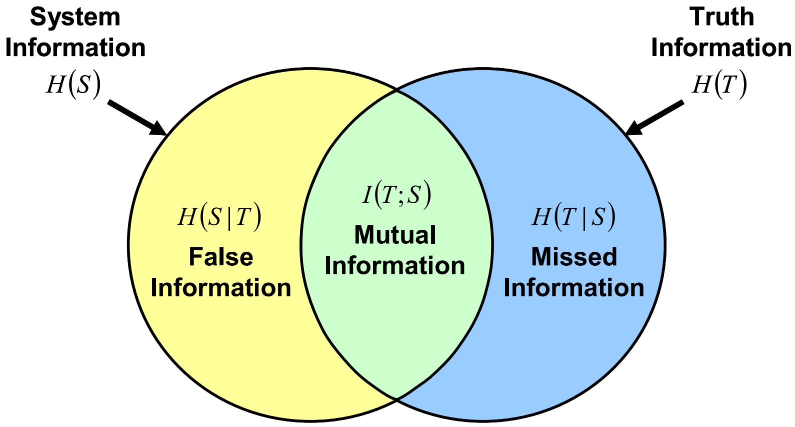


Figure 2. A joint entropy diagram [9] representing the five information measures relevant to performance evaluation.

For the recognition or classification problem, a confusion matrix, which represents the association of truth classes with system classes, fulfills the same role as the association matrix in the tracking problem; thereby, motivating the application of information theoretic metrics to the multi-class assignment problem.

Recall from Section 2 that the confusion matrix $D(t,s)$ can be normalized in order to represent the joint probability function $P_{t,s}$. The system and truth probability functions P_s and P_t, can be calculated by the normalized sums along rows or columns of $D(t,s)$, respectively. The conditional probability density functions $P_{s|t}$ and $P_{t|s}$ are simply $P_{t,s}/P_t$ and $P_{t,s}/P_s$. The information measures are easily computed from these probability terms as follows [8]:

$$H(T)=-\sum_{i=1}^{N}P_t(i)\log(P_t(i)), \tag{17}$$

$$H(S)=-\sum_{j=1}^{N}P_s(j)\log(P_s(j)), \tag{18}$$

$$H(S\mid T)=-\sum_{i,j=1}^{N}P_{t,s}(i,j)\log(P_{s|t}(j\mid i)), \tag{19}$$

$$H(T \mid S) = -\sum_{i,j=1}^{N} P_{t,s}(i,j)\log\left(P_{t|s}(i \mid j)\right), \tag{2}$$

$$I(T;S) = \sum_{i,j=1}^{N} P_{t,s}(i,j)\log\left(\frac{P_{t,s}(i,j)}{P_t(i)P_s(j)}\right). \tag{21}$$

Referring back to Figure 2, a perfect multi-class assignment system is one in which the false and missed information have been eliminated, thereby equating system information with truth information. From this reasoning, a single score can be calculated by summing the two conditional entropies $H(T \mid S)$ and $H(S \mid T)$. This score measures the amount of true information that a classifier missed plus the amount of false information that it generated. An optimal multi-class assignment system, from an information theoretic perspective, minimizes this score:

$$S_I = H(T \mid S) + H(S \mid T) = -\sum_{i,j=1}^{N} P_{t,s}(i,j)\log\left(P_{t|s}(i \mid j)\right) - \sum_{i,j=1}^{N} P_{t,s}(i,j)\log\left(P_{s|t}(j \mid i)\right). \tag{22}$$

When comparing this score to the metrics in Section 2, we can define the cost function as:

$$C_{t,s}(i,j)_{IC} = -\log P_{t|s}(i \mid j) - \log P_{s|t}(j \mid i). \tag{23}$$

Because comparisons between classifiers or algorithm parameters are best made with the same truth dataset, the entropy of the truth dataset can be used to normalize the score. In addition, the interrelation of the five information metrics can be used to reduce the scoring function to:

$$S_I \,/\, H(T) = 1 - f(T;S) + r(S \mid T), \tag{24}$$

where truth information completeness, $f(T;S)$, is a measure of the fraction of truth information collected by a classifier and false information ratio, $r(S \mid T)$, is the ratio of false information to truth information. These metrics are defined as:

$$f(T;S) = I(T;S) / H(T), \tag{25}$$

$$r(S \mid T) = H(S \mid T) / H(T). \tag{26}$$

We can graphically describe the overall performance of multiple classifiers by plotting the values for these two metrics and comparing their locations. The *information coverage plot* [8] for the classifiers introduced in Section 2 is shown in Figure 3. The normalized score $S_I / H(T)$ for classifiers A, B, C, and D is 0.855, 0.693, 0.481, and 0.666, respectively. This score represents the total amount of erroneous information, which is equivalent to the Manhattan distance from the point (0, 1) in the upper-left corner of the plot. The star indicates the location of a perfect classifier where no false information is generated and all truth information is captured. The dotted lines on the plot represent points in the information coverage space that have an equal amount of erroneous information.

The resulting rank order, using this evaluation method, from best to worst for the four classifiers is C, D, B, and A. We can see that the combination of these metrics is able to distinguish between classifiers A, B, and C, like the quadratic loss function mentioned above. We also see that these metrics are accounting for the spread of misclassification by ranking C better than B and B better than A, similar to the normalized error to information ratio. However, this method stands out from all the other metrics by declaring D to have better overall performance than B and A. This result may seem counter-intuitive, especially if the goal is to simply maximize accuracy, namely, the fraction of correct assignments. However, if the goal is to optimize the information content of the system output, this metric is a good choice. Indeed, classifier D contains less erroneous information in its confusion matrix than classifiers A and B, and contributes better information about the truth. Due to the fact that classifier D provides better information about the truth, one could argue that it has a higher overall performance than classifiers A and B.

The real value in the information theoretic metrics for evaluating classifier performance is that improvements are easier to make to classifiers that avoid distributing misclassification errors across the entire confusion matrix like classifier D. When the errors are contained to only a few cells, secondary classifiers could simply be activated when the initial classifier makes decisions where it is known to perform poorly.

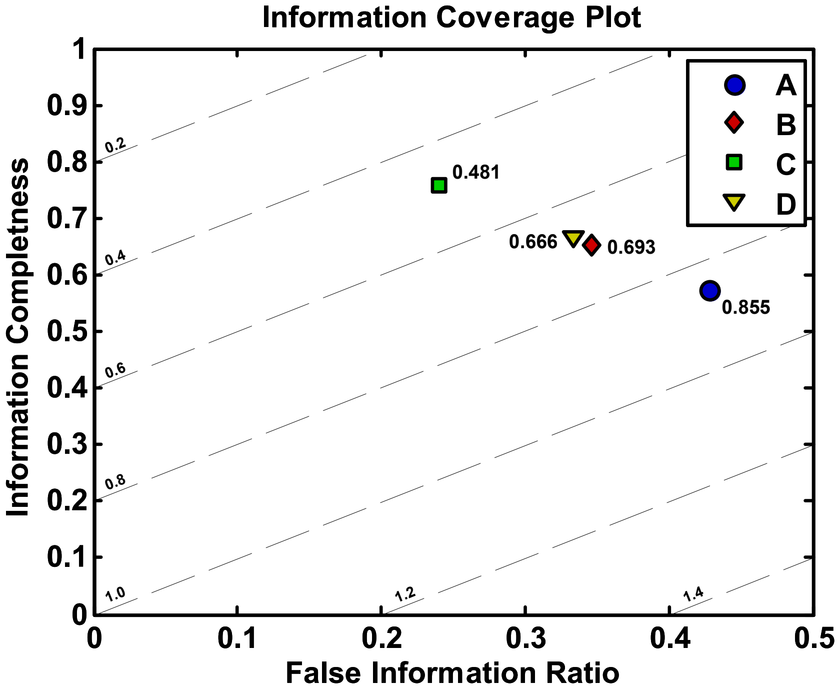
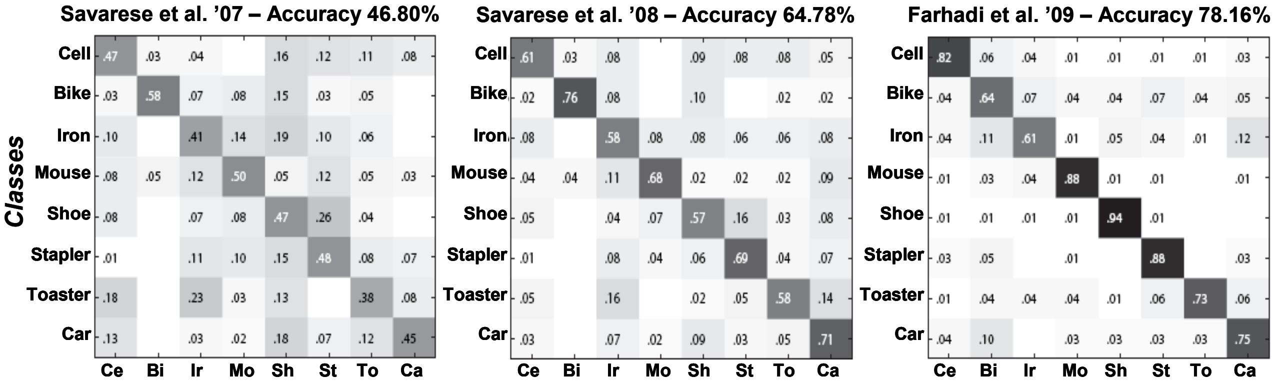


Figure 3. Information coverage plot showing the performance results of four classifiers. The values indicate the normalized score $S_I \,/\, H(T)$, which is equivalent to the amount of erroneous information.

4. APPLICATION

In this section, two examples are presented to demonstrate how the information theoretic metrics can be used to evaluate object recognition systems and classifiers. The first example draws on the results presented in a paper by Farhadi et al. [10][*], where three different object recognition algorithms were evaluated for eight different classes of common everyday objects. The confusion matrices for the three algorithms are shown in Figure 4, along with the percent accuracy for each classifier. By simply comparing the confusion matrices, it is difficult to describe a precise quantity of improvement from one system to another. While the percent accuracy does allow for such an improvement comparison, much of the information contained within in the confusion matrix is lost. Neither method provides a quantitative, comprehensive view of the system performance.



Savarese et al. '07 – Accuracy 46.80%

	Ce	Bi	Ir	Mo	Sh	St	To	Ca
Cell	.47	.03	.04		.16	.12	.11	.08
Bike	.03	.58	.07	.08	.15	.03	.05	
Iron	.10		.41	.14	.19	.10	.06	
Mouse	.08	.05	.12	.50	.05	.12	.05	.03
Shoe	.08		.07	.08	.47	.26	.04	
Stapler	.01		.11	.10	.15	.48	.08	.07
Toaster	.18		.23	.03	.13		.38	.08
Car	.13		.03	.02	.18	.07	.12	.45

Savarese et al. '08 – Accuracy 64.78%

	Ce	Bi	Ir	Mo	Sh	St	To	Ca
Cell	.61	.03	.08		.09	.08	.08	.05
Bike	.02	.76	.08		.10		.02	.02
Iron	.08		.58	.08	.08	.06	.06	.08
Mouse	.04	.04	.11	.68	.02	.02	.02	.09
Shoe	.05		.04	.07	.57	.16	.03	.08
Stapler	.01		.08	.04	.06	.69	.04	.07
Toaster	.05		.16		.02	.05	.58	.14
Car	.03		.07	.02	.09	.03	.05	.71

Farhadi et al. '09 – Accuracy 78.16%

	Ce	Bi	Ir	Mo	Sh	St	To	Ca
Cell	.82	.06	.04	.01	.01	.01	.01	.03
Bike	.04	.64	.07	.04	.04	.07	.04	.05
Iron	.04	.11	.61	.01	.05	.04	.01	.12
Mouse	.01	.03	.04	.88	.01	.01		.01
Shoe	.01	.01	.01	.01	.94	.01		
Stapler	.03	.05		.01		.88		.03
Toaster	.01	.04	.04	.04	.01	.06	.73	.06
Car	.04	.10		.03	.03	.03	.03	.75

Figure 4. Confusion matrices showing performance improvements over three object recognition systems presented in [10].

[*] This paper was selected for this example because it has a reasonable number of classes and a good description of the results.

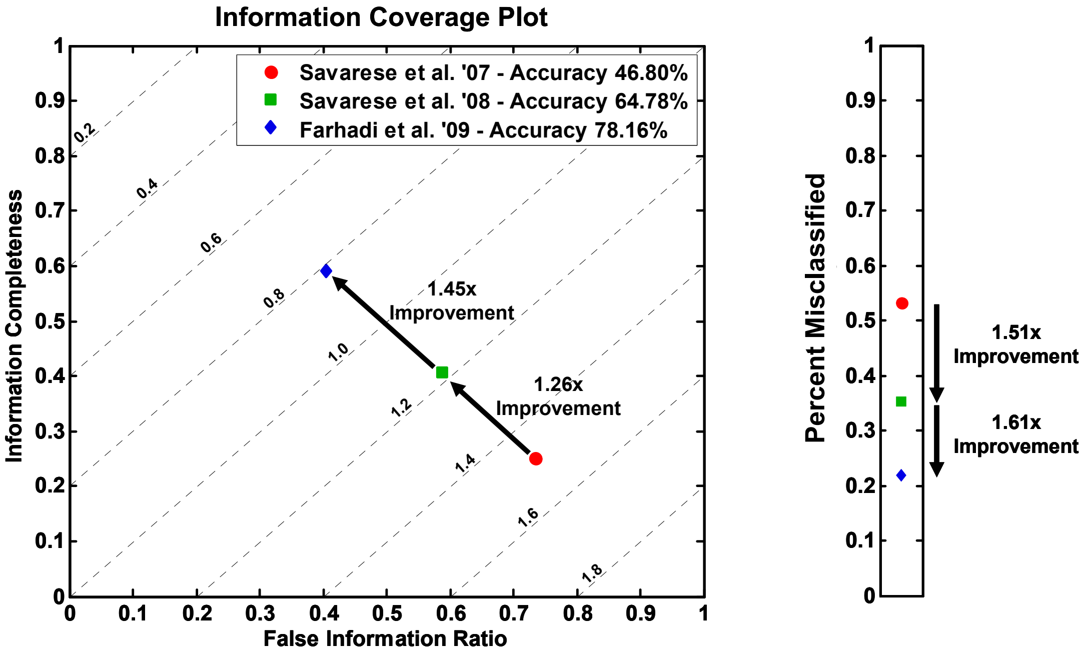


Figure 5. Information coverage plot and percent misclassified, or one minus percent accuracy, for each classifier. Both plots show performance improvement factors over the three years. The improvement factor for the ’09 classifier compared with the improvement factor of the ’08 classifier is more significant when using the information theoretic measures.

The information coverage plot offers a quantitative and comprehensive assessment. It captures the information of the entire confusion matrix and presents it in a two dimensional ROC-like metric space. Figure 5 shows the improvement of each classifier over the previous year in terms of the information theoretic metric and the percent accuracy. Improvement, here, is defined as the reduction of error. In the information coverage plot the error quantity is the total erroneous information as described in the previous section. For the accuracy plot, the error quantity is equivalent to one minus the percent accuracy, or the percent misclassified.

When each confusion matrix is presented as a data point in the metric space, as with the information coverage plot, one can see that the improvement factor from the ’08 to ’09 classifier compared to the improvement factor from the ’07 to ’08 classifier is more significant than previously realized by only looking at the percent misclassified. The ’09 paper not only reduces the fraction of misclassification, but the spread of the misclassification as well. As discussed in the previous section, a reduction in the misclassification spread increases the mutual information and decreases the conditional entropies between the truth and the declaration. In other words, the recognition system captures more information about the truth while conveying less false information. For comparison with the information theoretic metrics, Table 2 lists the metrics described in Section 2 for each of the classifiers.

More in-depth analysis can be performed on the data from the Farhadi paper [10] beyond just information completeness and the false information ratio. The conditional entropy of individual truth classes and decision classes can be evaluated to clearly identify where algorithms might be having problems. Similar to the overall conditional entropies $H(S \mid T)$ and $H(T \mid S)$, the conditional entropy by class are computed both on the system entropy given a truth class $H(S \mid T = i)$ and on the truth entropy given a system class $H(T \mid S = j)$. The former is a measure of correctness where as the latter is a measure of information completeness. Lower values indicate better performance. Figure 6 shows the results of three different classifiers, assuming equal number of samples over each truth class. The left plot shows the conditional entropy of each truth class for the three different classifiers. The ’08 algorithm performs the best at identifying bikes (class 2). Performance for the other classes is consistent with the overall classifier ranking. The right plot shows the conditional entropy of each decision class. The best overall classifier, the ’09 algorithm, yields the least amount of information when the bike class is declared, as indicated by the large uncertainty regarding the truth. This suggests that further improvements to this algorithm might be realized by addressing the issues with recognizing bikes.

Table 2. Comparison of metrics used to evaluate the performance of the three classifiers in Figure 4.

Evaluation Metric	Savarese '07	Savarese '08	Farhadi '09
Accuracy	46.8%	64.8%	78.2%
Kappa	0.389	0.593	0.756
TPR	0.466	0.643	0.786
FPR	7.64E-2	5.09E-2	3.06E-2
PPV	0.500	0.658	0.788
NPV	0.923	0.949	0.970
Rand index	0.866	0.911	0.947
F-score	0.475	0.647	0.784
Linear Loss	1.01	0.964	0.928
Quadratic Loss	0.888	0.846	0.813
Informational Loss	2.85	2.53	2.33
ξ	1.03	0.421	0.174
$S_I / H(T)$	1.484	1.180	0.814

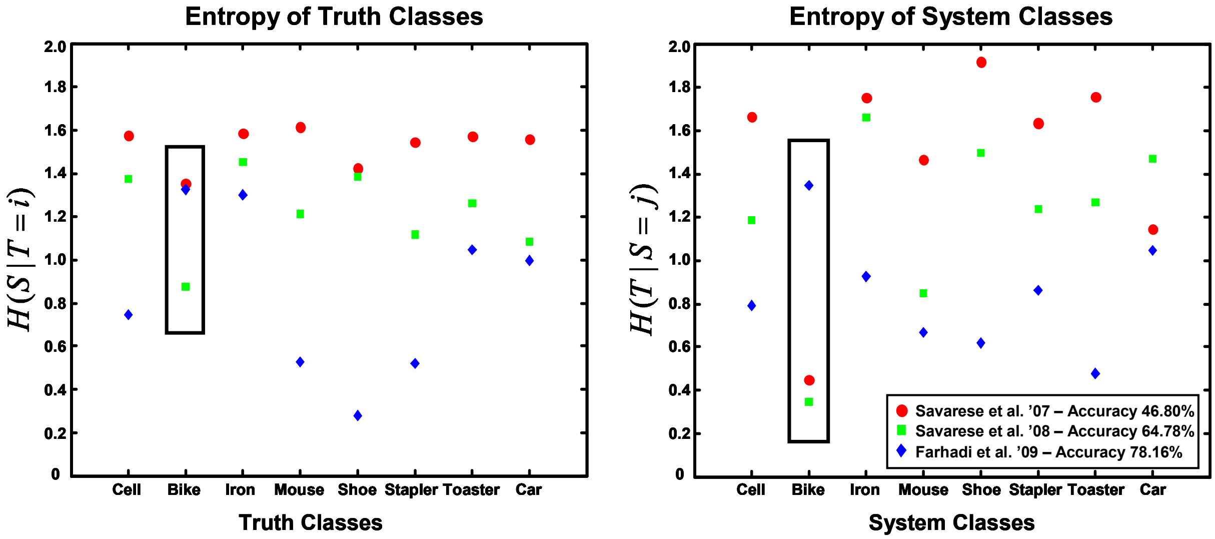
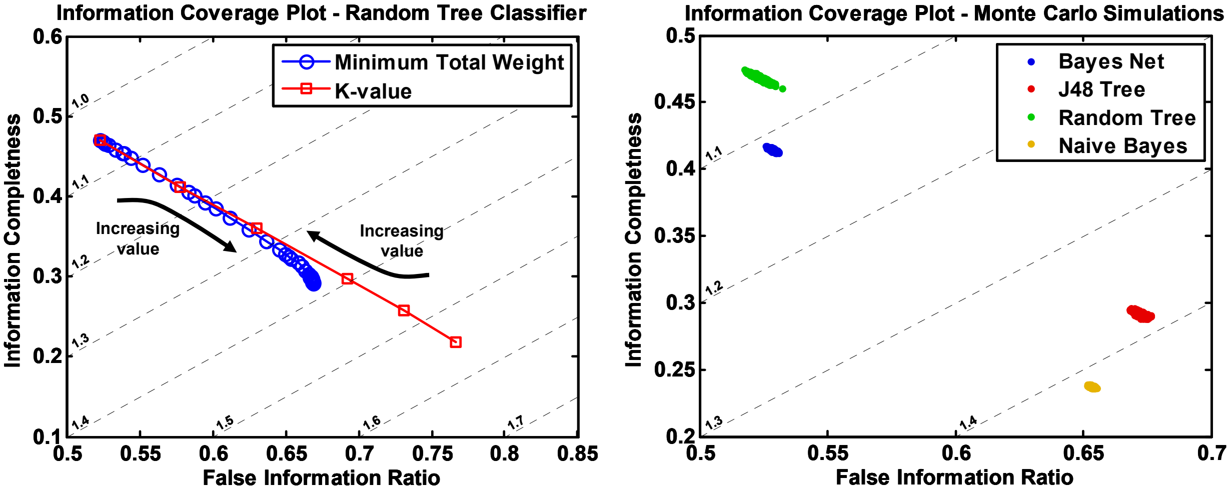


Figure 6. The conditional entropies by class. The system entropy conditioned on each truth class, the plot on the left, captures the correctness of the classifier on a certain truth class. The truth entropy conditioned on each system class, the plot on the right, captures how informative the classifier is when it declares a certain class. This analysis shows that the '09 classifier is challenged by the Bike class.

The second example demonstrates how the information coverage plot is used to tune classification algorithms as well as to describe the uncertainty in the performance evaluation through Monte Carlo simulations. For this example, common classification algorithms are used to classify the UCI Chess (King and Rook vs. King) dataset [11]. The Random Tree algorithm [6] has two parameters that can be adjusted: the minimum total weight for a leaf node in the tree and the K-value, which sets the number of randomly selected attributes to use when building the tree. By plotting $r(S\,|\,T)$ and $f(T;S)$ for different values of these parameters, the information coverage plot behaves like a ROC curve, describing the region of performance as a function of the parameters. Figure 7(a) shows the results of the Random Tree algorithm. The blue curve was generated by fixing the K-value at 6 and varying the minimum total weight from 1.0 to 20.0. For the red curve, the minimum total weight was fixed at 1.0 and the K-value was varied from 1 to 6. From this plot, we can see the algorithm has the best performance when the minimum total weight is set to 1.0 and the K-value is set to 6.

When evaluating the performance of a classifier, there is often uncertainty incurred from the data sampling used to test the classifier. A common approach used for lowering the uncertainty when testing an algorithm is to do *m-fold cross-validation* [12], which involves dividing the data into *m* disjoint sets of equal size. The classifier is then trained *m* times, each time a different set is withheld to be used for testing. The results are then summed together to yield a final confusion matrix. While this approach aids in giving a better estimate of the classifier performance, it is still lacking an indication of the spread of uncertainty on that estimate. Alternatively, a distribution on the evaluation performance results may be generated from Monte Carlo simulations, where the random seed used to generate the folds in cross-validation is varied. The results, plotted in information coverage space, give an estimate on the spread of uncertainty. Figure 7(b) shows the results of a thousand Monte Carlo runs for the Bayesian Network, J48 Tree, Random Tree, and Naïve Bayes algorithms [6]. The sample mean, standard deviation, and correlation for $r(S\,|\,T)$ and $f(T;S)$ are given in Table 3 for each classifier. When comparing two system configurations, a confidence score, or p-value, on the hypothesis that configuration A is better than configuration B can be computed as the percentage of runs where A achieves a lower overall erroneous information score than B. In this example, the p-value calculation is trivial since there is no overlap of the different distributions.

Both of the examples presented above demonstrated the usefulness of the information theoretic metrics. By displaying the truth information completeness and the false information ratio on the information coverage plot, we were able to provide a comprehensive evaluation for recognition systems and tune classifier parameters. Additional information theoretic measures using class-specific conditional entropies were used to pinpoint areas for improvement with the algorithms. Finally, we showed that the confidence of claiming one system's performance over that of another in information content can be assessed using Monte Carlo simulations.



(a) Parameter Tuning

(b) Error Estimate

Figure 7. Examples of how the information coverage plot is used to (a) assist with algorithmic parameter tuning and (b) describe the uncertainty in the evaluation estimate.

Table 3. The sample mean, standard deviation, and correlation for the false information ratio and the truth information completeness calculated from a thousand Monte Carlo simulations for four different classifiers.

Evaluation Metric	**Bayesian Network**	**J48 Tree**	**Random Tree**	**Naïve Bayes**
False Information Ratio (μ,σ)	0.528, 8.21E-4	0.672, 1.38E-3	0.523, 2.14E-3	0.653, 5.79E-4
Truth Information Completeness (μ,σ)	0.415, 7.82E-4	0.292, 1.35E-3	0.469, 2.05E-3	0.238, 4.99E-4
Correlation ρ	-0.831	-0.783	-0.947	-0.692

5. CONCLUSION

Existing metrics for multi-class assignment systems only capture partial and specific information on the performance. Attempts have been made to characterize system performance holistically using an ensemble of metrics or the full confusion matrix. However, such methods do not render an overall score and no agreed upon method exists to combine these metrics or to resolve conflicts between them. Information theory provides a unifying framework to evaluate the overall performance of multi-class assignment systems. The ROC-like information coverage plot offers an effective means by which to compare and visualize the overall performance between different classifiers. It also allows performance curves to be generated as system parameters are adjusted for optimal performance. Additional information theoretic measures can identify specific deficiencies in algorithm performance where additional algorithm development might be most productively applied. In the future, we intend to work on a more analytically rigorous estimation of the information theoretic measures from a confusion matrix and to determine the relationships between information theoretic measures and the Cramer-Rao lower bound.

REFERENCES

[1] Cohen, J., "A coefficient of agreement for nominal scales," Educational and Psychological Measurement 20(1), 37-46 (1960).

[2] Fawcett, T., "An introduction to ROC analysis," Pattern Recognition Letters 27, 861-874 (2006).

[3] Rand, W. M., "Objective criteria for the evaluation of clustering methods," Journal of the American Statistical Association 66, 846-850 (1971).

[4] van Rijsbergen, C. J., [Information Retrieval (2nd edition)], Butterworths, London, 112-138 (1979).

[5] Berger, J. O., [Statistical Decision Theory and Bayesian Analysis], Springer-Verlag, New York, 46-70 (1980).

[6] Witten, I. H., and Frank, E., [Data Mining: Practical Machine Learning Tools and Techniques (2nd edition)], Morgan Kaufmann, San Francisco, 143-185 (2005).

[7] Chen, H., Chen, G., Blasch, E., Douville, P., and Pham K., "Information theoretic measures for performance evaluation and comparison." 12th International Conference on Information Fusion, 874-881 (2009).

[8] Kao, E. K., Daggett, M. P., and Hurley, M. B., "An information theoretic approach for tracker performance evaluation," IEEE 12th International Conference on Computer Vision, 1523-1529 (2009).

[9] Cover, T. M., and Thomas, J. A., [Elements of Information Theory], Wiley-Interscience, New York, 20 (1991).

[10] Farhadi, A., Tabrizi, M, Endres, I., and Forsyth, D., "A Latent Model of Discriminative Aspect," IEEE 12th International Conference on Computer Vision, 948-955 (2009).

[11] Blake, C. L., and Merz, C. J., "UCI repository of machine learning databases," Univ. of California at Irvine, Department of Information and Computer Sciences, http://archive.ics.uci.edu/ml/datasets.html (1998).

[12] Duda, R. O., Hart, P. E., and Stork, D. G., [Pattern Classification (2nd edition)], Wiley-Interscience, New York, 483-485 (2001).