METHOD ARTICLE

REVISED Formal definition of the MARS method for quantifying the unique target class discoveries of selected machine classifiers [version 2; peer review: 2 approved]

Felipe Restrepo[1], Namrata Mali[2], Alan Abrahams[3], Peter Ractham[4]

[1]Department of Industrial and Systems Engineering Information,, Virginia Tech, Virgina, 24061, USA
[2]Department of Computer Science, Virginia Tech, Virginia, 24061, USA
[3]Department of Business Information Technology, Virginia Tech, Virginia, 24061, USA
[4]Department of Management Information Systems, Thammasat University, Bangkok, 10200, Thailand

v2 **First published:** 04 Apr 2022, **11**:391
https://doi.org/10.12688/f1000research.110567.1
Latest published: 01 Jul 2022, **11**:391
https://doi.org/10.12688/f1000research.110567.2

Abstract

Conventional binary classification performance metrics evaluate either general measures (accuracy, F score) or specific aspects (precision, recall) of a model's classifying ability. As such, these metrics, derived from the model's confusion matrix, provide crucial insight regarding classifier-data interactions. However, modern- day computational capabilities have allowed for the creation of increasingly complex models that share nearly identical classification performance. While traditional performance metrics remain as essential indicators of a classifier's individual capabilities, their ability to differentiate between models is limited. In this paper, we present the methodology for MARS (Method for Assessing Relative Sensitivity/ Specificity) ShineThrough and MARS Occlusion scores, two novel binary classification performance metrics, designed to quantify the distinctiveness of a classifier's predictive successes and failures, relative to alternative classifiers. Being able to quantitatively express classifier uniqueness adds a novel classifier-classifier layer to the process of model evaluation and could improve ensemble model-selection decision making. By calculating both conventional performance measures, and proposed MARS metrics for a simple classifier prediction dataset, we demonstrate that the proposed metrics' informational strengths synergize well with those of traditional metrics, delivering insight complementary to that of conventional metrics.

Keywords

Machine learning, Binary classification, Classifier performance evaluation, Classifier selection optimization, Classifier comparative uniqueness

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 01 Jul 2022	✓ view	✓ view
version 1 04 Apr 2022	? view	? view

1. **Timothy A. Warner**, West Virginia University, Morgantown, USA

2. **Samir Chatterjee**, Claremont Graduate University, Claremont, USA

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Artificial Intelligence and Machine Learning gateway.

Corresponding author: Peter Ractham (peter@tbs.tu.ac.th)

Author roles: Restrepo F: Data Curation, Formal Analysis, Validation, Visualization, Writing – Original Draft Preparation; **Mali N**: Data Curation, Formal Analysis, Validation, Visualization, Writing – Original Draft Preparation; **Abrahams A**: Conceptualization, Funding Acquisition, Methodology, Supervision, Writing – Review & Editing; **Ractham P**: Project Administration, Supervision, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

Grant information: This study was supported by grants from the Virginia Tech Data & Decision Sciences (D&DS) and Virginia Tech Institute for Society, Culture, and the Environment (ISCE): Principal Investigators: Alan S. Abrahams and Laura P. Sands.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Restrepo F, Mali N, Abrahams A and Ractham P. **Formal definition of the MARS method for quantifying the unique target class discoveries of selected machine classifiers [version 2; peer review: 2 approved]** F1000Research 2022, **11**:391 https://doi.org/10.12688/f1000research.110567.2

First published: 04 Apr 2022, **11**:391 https://doi.org/10.12688/f1000research.110567.1

REVISED **Amendments from Version 1**

We incorporated a new figure, MARS ShineThrough bar chart (MARS charts section), which allows for the prompt visualization of the classifiers' individual ETP but provides no information about combined classifier target-class discovery efforts. The discussion section was extended to explain the circumstances under which MARS metric usage is ideal and to further emphasize that for applications with different tradeoffs, an inverted MARS evaluation method, aimed at maximizing true negatives, may be preferable.

Any further responses from the reviewers can be found at the end of the article

Introduction

Traditionally, binary classification performance has been assessed using a combination of statistical measures derived from the classifier's confusion matrix (accuracy, precision, recall/sensitivity, specificity, F score), or the classifier's various confusion matrices, in the case of classifications at different cut-off thresholds (ROC curve, AUC metric). Accuracy is defined as the percentage of correct predictions out of all predictions. Precision is the percentage of predicted positives that are true. Recall (sensitivity) is the percentage of actual positives that are correctly predicted. Specificity is the percentage of actual negatives that are correctly predicted. F scores (various variants like F_1, F_2) combine precision and recall, weighting each equally, or unequally, to account for different misclassification costs. Finally, for binary classifiers that assign a probability or score to predictions, ROC curves and AUC metrics account for these ranked predictions, allowing for sensitivity and specificity to be observed at different cut-off thresholds. To plot the ROC curve and assess AUC, sensitivity and specificity are measured @k, where k is the number of top-ranked predictions and increases from 1 to the total number of observations in the dataset. Effective classifiers demonstrate a "bulge" in the ROC curves, and concomitant AUC close to 1, indicating that they discover far more true positives in the top-ranked k items, than would be expected in a random selection of k items. Notably, none of these conventional metrics assess the distinctiveness (uniqueness) of the classifier's predictions, relative to other classifiers. In other words, conventional metrics are unable to assess what percentage of true positives ('hits') are found only by the current algorithm but not by alternatives, nor what percentage of false negatives ('misses') were missed by the current algorithm but not by alternatives.

Prior to modern-day computational capabilities, the inability to quantify classifier uniqueness had not been seen as a significant limitation, as available computing power did not allow for the use of big-data or complex classifiers, resulting in a low-diversity classifier prediction sample space for most applications. However, within the context of modern-day computational power, which allows for the use of high-volume data to train complex ML classifiers for tasks beyond traditional classification/regression, e.g., discovery-driven tasks, such as flagging potentially hazardous products via online reviews (discussed below); the inability to quantify how many, and what proportion, of a classifier's correct (and incorrect) predictions are exclusive to that classifier is a significant limitation. Especially considering that complex models may often report equal accuracy (or precision, or recall, or AUC), but have fundamentally different decision boundaries, resulting in a high-diversity prediction sample space – hence, the classifiers may each have the unique ability to identify distinct observations from the target class, and this classifier uniqueness ought to be assessable.

Such assessments about classifier uniqueness have been made possible through the use of novel MARS (Method for Assessing Relative Sensitivity) ShineThrough and MARS Occlusions scores, whose software-level implementation was recently described in Ref. 19. However, since[19] focuses solely on the usage and interpretation of the software artifact's outputs, it does not outline the methodological framework used to generate ShineThrough and Occlusion scores. Thus, in this paper, we present the mathematical foundations behind MARS metrics and their corresponding software artifact. Furthermore, we also provide step-by-step sample calculations that illustrate the inner workings of Shinethrough and Occlusion scores for a simple dataset. Being able to quantitatively assess classifier uniqueness has multiple benefits: better decisions could be made about combining complementary classifiers (vs duplicative classifiers), and improved characterizations could be run of where particular classifiers 'shine through' (spot true positives that no other classifiers spot) or 'occlude' (hide or miss observations in the target class, by mistakenly classifying those observations as false negatives, when all other classifiers were able to spot those observations as true positives).

As an example of the problematic omission of exclusivity metrics in the evaluation and comparison of classifiers, consider the following cases. Recently,[1] evaluated the generalized, binary predictive ability of eight classifiers across ten datasets. ROC curve values for the top-ranked classifiers revealed that Support Vector Machine (SVM), Artificial Neural Network (ANN), and Partial Least Squares Regression (PLS) classifier performances were nearly identical across all datasets.[2] compared the performance of several classifiers, namely, Random Forest (RF), Decision Tree (DT), and k-nearest neighbors (kNN), using binary classification schemes for variable stars. Similar to Refs. 1,2's precision, recall,

and F_1 scores indicated that all three classifiers performed nearly identically.[3–5] reported similar outcomes, with virtually equal performance metric values across the top n-ranked classifiers. In all these cases, while the performance of the classifiers is nearly identical according to conventional classifier evaluation metrics, the classifiers clearly made different false positive and false negative errors, and thus triumphed, or failed, relative to other classifiers on particular observations. Clearly, the scope of traditional statistical performance measures is too narrow to provide the insight required to distinguish between the top n-ranked classifiers based on their respective exclusive hits or misses. Novel classifier exclusivity metrics are needed to illustrate the success or failure of classifiers on particular observations, relative to their competing classifiers. These exclusivity metrics should reflect the extent to which a classifier exclusively finds ("shines through") observations in the target class (that are not spotted by competing classifiers), or exclusively misses ("occludes") observations in the target class that are spotted by competing classifiers.

Consider a classification task where the data scientist is attempting to identify safety concerns expressed by consumers in millions of online product reviews (e.g., see Refs. 6–9), using alternative candidate classifiers C_1 and C_2. The classification task is critical: missed safety concerns are unaddressed product hazards that could injure current or future product users. Assume the two competing classifiers, C_1 and C_2, both have precision of 80%, and recall of 80%, superficially (i.e., prima facie) indicating the classifiers have similar performance. However, if we are able to take into consideration the exclusivity of the classifier's predictions ("shine through" and "occlusion"), we may find that C_1 finds a significant proportion of the target class (safety concerns, in this observation) that C_2 misses ("occludes"). Assessing classifier exclusivity is thus essential to revealing that two classifiers with 80% precision are by no means identical in their target- observation discovery ability, and may be complementary, rather than simply competing. This realization allows the data scientist to discover more safety concerns, through intelligent classifier combination (e.g., taking true positives from both classifiers), rather than the data scientist simply deciding to eliminate a superficially comparable classifier (when regarding conventional classifier performance metrics only prima facie).

Hence, while traditional performance metrics are highly efficient at identifying elite models, they tend to fall short when the task at hand requires that these (elite) models be differentiated, particularly so if the source data is of high volume.

In this paper, we present the methodology for MARS ("Method for Assessing Relative Sensitivity"), a novel approach that evaluates the comparative uniqueness of a classifier's predictions, relative to other classifiers.[19] By mathematically defining MARS 'ShineThrough' and 'Occlusion' scores, we demonstrate how these metrics assess model performance as a function of the model's ability to exclusively capture unique true positives not found by the other classifiers ('ShineThrough') and the model's inability to capture true positives found by the other classifiers ('Occlusion'). These metrics, designed to complement widely used traditional and alternative measures, add another layer to classifier assessment, provide crucial insight that helps better distinguish and explain the behavior of the top n-ranked classifiers, and can be further extended to find optimal complementary classifier combinations for target-class discovery.

Related work

Binary classification Machine Learning (ML) performance metrics provide quantitative insight pertaining to different facets of a classifier's true behavior, i.e., its performance on unseen data. For example, while precision is defined as the proportion of predicted positives that are actually positives, recall (sensitivity) is the overall proportion of actual positives that were correctly labelled as such.[10] These metrics, derived from the classifier's confusion matrix (Figure 1), offer

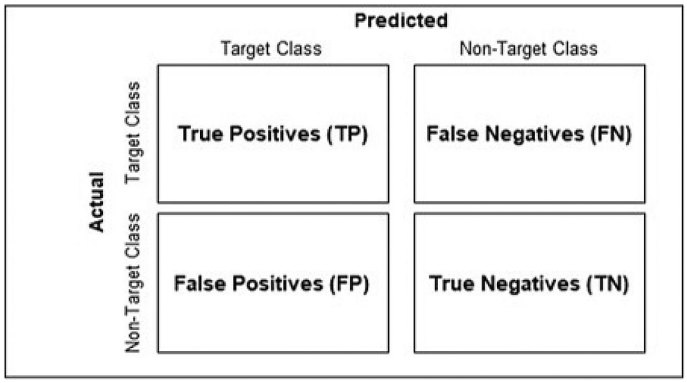

Figure 1. Format of a conventional classifier confusion matrix.

complementary assessments concerning the classifier's ability to detect and correctly label true positives, as evidenced by their mathematical definitions:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Abbreviations used: TP = True Positives, FP = False Positives.

$$\text{Recall}\,(\text{Sensitivity}) = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Abbreviations used: FN = False Negatives.

Similar to sensitivity, which calculates the model's true positive rate, specificity evaluates the overall proportion of negatives that were correctly labelled by the classifier (true negative rate).[11] Consequently, it follows a similar formulation:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Abbreviations used: TN = True Negatives.

These metrics (precision, recall, specificity) provide crucial insight relating to classifier-class interactions. Other measures, such as accuracy and F score,[12] provide a more generalized interpretation of model behavior. F score, defined as the harmonic mean of precision and recall, evaluates the classifier's performance across three confusion matrix components: TP, FP, FN, and can be defined as follows:

$$\text{F}_\beta = \left(1 + \beta^2\right) \cdot \frac{\text{precision} \cdot \text{recall}}{\left(\beta^2 \cdot \text{precision}\right) + \text{recall}}$$

Where β is arbitrarily chosen such that recall is β times as important as precision. The two most commonly used implementations are F_1 and F_2 scores.[13–15]

Overall accuracy, unlike the aforementioned metrics, incorporates all four confusion matrix components into its calculations:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

As for visual metrics and evaluation of a classifier over multiple classification cut-off thresholds (ranked predictions), Receiver Operating Characteristics (ROC) curves[16,17] and Precision-Recall (PR) curves are generally considered to be the standard. ROC curves display what proportion of the total target class items were found by the classifier (sensitivity) in the x top- ranked target class predictions (x-axis).

Precision-Recall [PR] curves are sometimes used as an alternative to ROC curves,[18] to illustrate fluctuations in hit- and miss-rates, as increasing numbers of top-ranked observations are considered by a classifier. Notably, neither ROC curve nor PR curves indicate how many of the true positives in the top-ranked predictions are exclusive to the current classifier (i.e., were target-class items not found by any other classifier), nor how many of the false negatives are exclusive to the current classifier (i.e., were target-class items correctly found by all the other classifiers). Regarding this, the use of the MARS software artifact, proposed in Ref. 19, has been suggested as a way to overcome this limitation, which we further validate in this paper by presenting the mathematical foundations behind the software-level implementation of the MARS metrics.

Methods

We assess overall classifier uniqueness across two separate dimensions: MARS ShineThrough and MARS Occlusion scores. These performance measures are briefly defined in Ref. 19 as:

1. MARS ShineThrough Score: The proportion of exclusive true positives discovered only by the classifier under consideration, relative to the total number of unique true positives (i.e., counting each target-class observation once only, if it is found by any classifier) discovered across all classifiers.

2. MARS Occlusion Score: The classifier's proportion of exclusive false negatives (missed only by the current classifier) that were correctly labelled by all the other classifiers relative to the total number of unique true positives discovered across all classifiers (i.e., counting each target-class observation once only, if it is found by any classifier).

These performance measures are rigorously analyzed and mathematically anatomized in the subsections *MARS Shinethrough scores* and *MARS Occlusion scores* below. Note that the approach described in the following sections can be easily adapted to true negatives and false positives, instead of true positives and false negatives, but is omitted for brevity (as the calculations are identical).

Notation reference

Table 1 provides a quick-reference glossary of the symbols used in our definitions.

MARS ShineThrough Scores

Let n be the number of observations in a given dataset and J the set of classifiers, under consideration. Similarly, let y_i be classifier's predicted class label and t_i the true class label (0 or 1) at observation i.

Then, we can define the total number of true positives (TTP_{all}) as the sum, over n observations, of the maximum value of the product between predicted and true class labels across all j classifiers:

$$\mathrm{TTP}_{\mathrm{all}} = \sum_{i}^{n} \max\left(y_{i,\mathrm{C}_j} \cdot t_i, \forall \mathrm{C}_j \in J\right) \tag{1}$$

To determine the total number of exclusive true positives (ETP_{C_w}) discovered by the classifier of interest, C_w,j, i.e., target class observations found only by the current classifier and not found by the other classifiers, we use:

$$\mathrm{ETP}_{\mathrm{C}_w} = \sum_{i}^{n} \left(y_{i,\mathrm{C}_w} \cdot t_i\right) - \max\left(y_{i,\mathrm{C}_j} \cdot t_i, \forall \in J \neq \mathrm{C}_w\right) \cdot Zi \tag{2}$$

Where we sum (over n observations) the difference between the product of predicted and actual class labels and the maximum value of the same product across the remaining j -1 classifiers. Additionally, we multiply the latter by constant Z_i, defined as:

Table 1. Glossary of symbols used.

Symbol	Definition
i	Observation number
j	Classifier number
n	Total number of observations
y_{i,C_j}	Predicted class label for observation i, predicted by classifier j
t_i	True class label for observation i
J	Set of classifiers
C_w	Classifier of interest
C_j	Classifier j
Z_i	Constant defined in (2.1) for observation i
R_i	Constant defined in (4.1) for observation i
TTP_{all}	Total number of unique true positives across all classifiers
ETP_{C_j}	Exclusive true positives found by classifier j
EFN_{C_j}	Exclusive false negatives for classifier j

$$Z_i = \begin{cases} 1 \Longleftrightarrow y_{i,C_w} = 1, & t_i = 1 \\ 0, & otherwise \end{cases} \tag{2.1}$$

Consequently, the sum at observation i will have a non-zero value if and only if the classifier's predicted and actual labels belong to the target class.

Then, using (1) and (2), we calculate the ShineThrough Score for classifier j as follows:

$$\text{ShineThrough}_{C_j} = \frac{\text{ETP}_{C_w}}{\text{TTP}_{\text{all}}} \tag{3}$$

Hence, MARS ShineThrough provides a much-needed numerical interpretation of the classifier's comparative uniqueness, i.e., what proportion of the total number of true positives were exclusively identified by the classifier under consideration, relative to the competing classifiers. Occlusion scores, on the other hand, provide insight relating to the classifier's comparative weaknesses.

MARS occlusion scores

We define the total number of expected false negatives (EFN_{C_w}) labelled by the classifier of interest, C_w, and correctly labelled by all of the remaining

$j - 1$ classifiers as:

$$\text{EFN}_{C_w} = \sum_{i}^{n} \min\left(y_{i,j,C_j} \cdot t_i, \forall \in J \neq C_w\right) \cdot R_i \tag{4}$$

Where we find the minimum value of $y_{i,j} \cdot t_i$ across the remaining $j - 1$ classifiers and multiply the output by binary constant R_i, defined as:

$$R_i = \begin{cases} 1 \Longleftrightarrow y_{i,C_w} = 0, & t_i = 1 \\ 0, & otherwise \end{cases} \tag{4.1}$$

Thus, the summation will have a non-zero value at observation i if and only if the classifier under consideration incorrectly labelled the target class. Using (1) and (4), we then define the MARS Occlusion score for C_w as:

$$\text{Occlusion}_{C_w} = \frac{\text{EFN}_{C_w}}{\text{TTP}_{\text{all}}} \tag{5}$$

Where we divide EFN_{C_w} by TTP_{all} to determine what proportion of the classifier's false negatives are true positives for the remaining $j - 1$ classifiers, therefore, quantitatively assessing the classifier's comparative weaknesses.

Use cases

For the purposes of illustration, in the following subsections, we provide a stylized dataset and step-by-step, worked examples showing the computation of the MARS ShineThrough and MARS Occlusion scores, as well as the plotting of multiple MARS scores visually, in MARS charts.[20]

While we provide an arbitrary, stylized dataset in this paper (to facilitate the understanding of the step-by-step examples), MARS metric performance on a real dataset can be found in Ref. 19. However, the latter does not provide any worked-out examples or rigorous mathematical explanations beyond the software-artifact's outputs.

Dataset

We created a simple, binary classification dataset with ten observations, each assigned an artificially generated "true" class label, for illustrative purposes. We also generated (predicted) labels for arbitrary classifiers: $J = \{C_1, C_2, C_3, C_4\}$. Actual (true) and classifier (predicted) labels can be seen in Table 2.

MARS ShineThrough score metric: example computation

In order to calculate MARS scores, we first determine the total number of true positives discovered across all four classifiers using Eq. (1), that is:

Table 2. Sample classifier prediction matrix.

		Observation ID, for Observation *i*									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Predicted	C_1	1	0	0	0	1	1	1	1	0	0
class (C_{1-4})	C_2	1	1	1	1	0	0	0	0	1	0
	C_3	0	1	0	0	1	0	0	0	1	0
	C_4	0	1	1	1	0	0	1	0	0	1
	Actual class	0	1	0	1	0	1	1	1	0	1

$$\mathrm{TTP}_{\mathrm{all}} = \sum_{i=1}^{10} \max\left(y_{i,\mathrm{C}_j} \cdot t_i, \forall \mathrm{C}_j \in J\right)$$

We illustrate the sum's inner calculations for the first two observations below:

$$@\, i = 1, \text{true class} = 0:$$

$$\max\left(1 \times 0, 1 \times 0, 0 \times 0, 0 \times 0\right) = 0$$

$$@\, i = 2, \text{true class} = 1:$$

$$\max\left(0 \times 1, 1 \times 1, 1 \times 1, 1 \times 1\right) = 1$$

Thus, the sum at $i = 10$ would be:

$$\mathrm{TTP}_{\mathrm{all}} = \sum_{=1}^{10} 0+1+0+1+0+1+1+1+0+1 = 6$$

Summing over all ten observations yields the value of 6, indicating that every target-class observation was correctly labelled by at least one classifier. This can be double-checked by looking at the classifiers' target class predictions in Table 2 (i = 2,4,6,7,8,10).

To calculate individual ShineThrough scores for the classifier under consideration, we divide the total number of exclusive true positives found by C_w by the total number of unique true positives (i.e., correctly classified observations in the target-class) across all classifiers (Eq. (3)). We demonstrate the ETP calculation procedure for C_1 in Table 3.

Table 3. Sample ShineThrough calculations for C_1. Z_i, constant defined for observation *i*.

Observation (*i*)	**Pred. class (y_i)**	**True class (t_i)**	**Z_i**	**Inner sum - Eq. (2)**
1	1	0	0	$(1 \times 0) - \max(1 \times 0, 0 \times 0, 0 \times 0) \times 0$ = **0**
2	0	1	0	$(0 \times 1) - \max(1 \times 1, 1 \times 1, 1 \times 1) \times 0$ = **0**
3	0	0	0	$(0 \times 0) - \max(1 \times 0, 0 \times 0, 1 \times 0) \times 0$ = **0**
4	0	1	0	$(0 \times 1) - \max(1 \times 1, 0 \times 1, 1 \times 1) \times 0$ = **0**
5	1	0	0	$(1 \times 0) - \max(0 \times 0, 1 \times 0, 0 \times 0) \times 0$ = **0**
6	1	1	1	$(1 \times 1) - \max(0 \times 1, 0 \times 1, 0 \times 1) \times 1$ = **1**
7	1	1	1	$(1 \times 1) - \max(0 \times 1, 0 \times 1, 1 \times 1) \times 1$ = **0**
8	1	1	1	$(1 \times 1) - \max(0 \times 1, 0 \times 1, 0 \times 1) \times 1$ = **1**
9	0	0	0	$(0 \times 0) - \max(1 \times 0, 1 \times 0, 0 \times 1) \times 0$ = **0**
10	0	1	0	$(0 \times 1) - \max(0 \times 1, 0 \times 1, 1 \times 1) \times 0$ = **0**

Finally, we use Eq. (3) to obtain C_1 ShineThrough scores:

$$\text{ShineThrough}_{C_1} = \frac{2}{6}$$

This reveals that C_1 alone accounts for one third of the discovered target class observations, suggesting its behavior is fairly unique amongst its peers. The calculations can be easily verified by looking at observations $i = 6$ and $i = 8$ in Table 2. Additionally, we can also calculate combined ShineThrough scores for two or more classifiers by summing the number of unique TPs discovered by the models, i.e., their combined ETP.

For example, using Table 2, we can obtain the combined ShineThrough score for C_1 and C_4 using Eq. (1), (2), and (3), as follows:

$$\text{ETP}_{C_{1,4}} = \sum_{i=1}^{10} \left(y_{i,C_{1,4}} \cdot t_i \right) - \max \left(y_{i,C_j} \cdot t_i, \forall \in J \neq C_{1,4} \right) \cdot Z_i$$

$$@\ i = 10:$$

$$\text{ETP}_{C_{1,4}} = \sum_{i=1}^{10} 0+0+0+0+0+1+1+1+0+1 = 4$$

$$\text{ShineThrough}_{C_{1,4}} = \frac{4}{6} = \frac{2}{3}$$

This combined-ShineThrough indicates that two-thirds of the total target class observations Eq. (6), were exclusively discovered by classifiers C_1 and C_4, revealing that when combined, the classifiers are highly capable of target-class discovery, relative to the remaining classifiers. Note that originally (prior to combining classifiers), the observation at $i = 7$ was not considered to be exclusive for any of the classifiers, however, once C_1 and C_4 had their predictions combined, it became exclusive for $C_{1,4}$.

MARS occlusion score metric: example computation

As for occlusions scores, we can calculate the total number of exclusive false negatives (missed only by the current classifier) that were correctly classified by the other classifiers following Eq. (4):

$$\text{EFN}_{C_w} = \sum_{i}^{n} \min \left(y_{i,j,C_j} \cdot t_i, \forall \in J \neq C_w \right) \cdot R_i$$

In the case of C_1, the first two iterations of the sum are as follows:

$$@\ i = 1:$$

$$y = 1,\ t_1 = 0,\ R_i = 0$$

$$\min\left(1 \times 0, 0 \times 0, 0 \times 0\right) \times 0 = 0$$

$$@\ i = 2:$$

$$y = 0,\ t_2 = 1,\ R_i = 1:$$

$$\min\left(1 \times 1, 1 \times 1, 1 \times 1\right) \times 1 = 1$$

Following the same procedure, the final sum at $i = 10$ would be:

$$\text{EFN}_{C_1} = \sum_{i=1}^{10} 0+1+0+0+0+0+0+0+0+0 = 1$$

Then, we calculate the Occlusion score for classifier C_1 using Eq. (5):

$$\text{Occlusion}_{C_1} = \frac{\text{EFN}_{C_1}}{\text{TTP}_{\text{all}}} = \frac{1}{6}$$

Unlike ShineThrough scores (where higher scores suggest better performance), with Occlusion scores it is the case that lower scores suggest better performance. In the case of C_1, its Occlusion score reveals that 16% of the target class observations discovered by all other competing classifiers are being misclassified by C_1. Similar to ShineThrough scores, we can also sum classifier exclusive FN predictions to calculate combined Occlusion scores. For example, for C_1 and C_3, whose combined predictions only have false negatives correctly labelled by the other classifiers (C_2 or C_4) at observation $i = 4$ (Table 1), we can calculate combined $\text{Occlusion}_{1,3}$ as follows:

$$@\, i = 4:$$

$$y = 0,\ t_4 = 1,\ R_i = 1$$

$$\min\left(1 \times 1, 1 \times 1\right) \times 1 = 1$$

Then,

$$\text{Occlusion}_{C_{1,3}} = \frac{1}{6}$$

Occlusion scores for the combined classifier, $C_{3,4}$, indicate that one third of the target class labels were misclassified by the combination of classifier C_3 and classifier C_4, but correctly labelled by at least one of the remaining $j - 1$ classifiers.

MARS charts

MARS ShineThrough and Occlusion scores can also be visualized, allowing for the rapid depiction of the classifiers' relative uniqueness. For our example dataset and classifiers above, the MARS metrics can be transformed from proportions (of total true positives) to counts (of unique hits or misses), and visualized, across individual and combined classifiers, as seen in Figures 2-4, using a bubble-chart style format. Figure 2 is the MARS ShineThrough chart for classifiers $C_{1\text{-}4}$; the radius of the yellow circle represents the number (count) of exclusive true positives found by the classifier on the y-axis. The radius of the orange circle represents the number of exclusive true positives found by both the classifier on the y-axis and x-axis, i.e., combined ShineThrough. Figure 3 is the MARS Occlusion chart: the radius of the red circle represents the classifier of interest (y-axis) number of false negatives (correctly labelled by the other classifiers) and the radius of the orange circle represents the combined number of exclusive false negatives labelled by the classifiers on the x and y-axis (correctly labelled by the remaining classifiers).

Note that orange circles can only be as small as their respective yellow or red counterparts, which in turn may be as small as zero (indicating that the classifier found no exclusive true positives or false negatives).

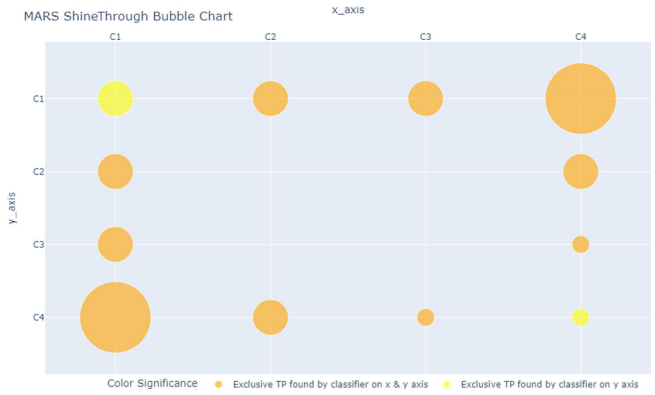

Figure 2. MARS ShineThrough Chart, comparing count (represented by bubble radius) of target-class observations (True Positives) exclusively spotted by classifiers C1 and the pairwise classifier combinations. Bubble size is proportional to ShineThrough score: the larger the bubble, the higher the classifier(s) ShineThrough score.

Figure 3. MARS Occlusion Chart, comparing count (represented by bubble radius) of target-class observations (False Negatives) exclusively missed by classifiers C1-4 and the pairwise classifier combinations. Bubble size is proportional to Occlusion score: the larger the bubble, the higher the classifier(s) Occlusion score.

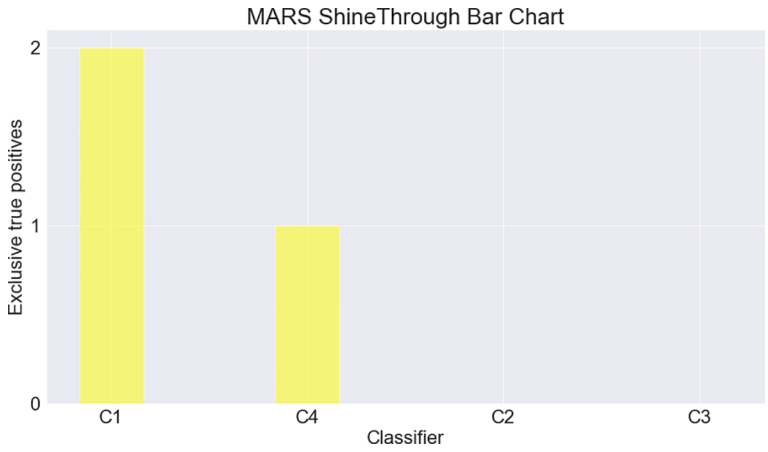

Figure 4. MARS ShineThrough Bar Chart, comparing count of target-class observations exclusively found by classifiers C1-4.

Individual classifier ETP counts can also be displayed via bar chart (Figure 4), allowing for prompt visual analysis of the classifiers' individual capabilities, but providing no information about combined classifier target-class discovery efforts.

Discussion

Conventional metrics (Table 4; columns 2-4) immediately identify C_4 as the unquestionably strongest classifier, due to its high accuracy (column 2), precision (column 3), and recall (column 4) values. However, notice that the information presented in these columns (2-4) does not go beyond identifying the individually strongest classifier, there is no insight

Table 4. Traditional vs MARS Metrics for the worked example.

	Metrics				
Classifier	**Accuracy**	**Precision**	**Recall**	**ST**	**OCC**
C_1	0.50	0.60	0.50	**0.33**	**0.16**
C_2	0.20	0.40	0.33	0.0	0.0
C_3	0.30	0.33	0.16	0.0	0.0
C_4	**0.70**	**0.80**	**0.66**	0.16	0.0

For brevity, we show only arbitrary selected classifier combinations here, rather than all possible classifier combinations. The best performing individual, and combined, classifier, on each metric, is shown with cell **bolded**.

relating to the classifiers' decision boundaries or prediction uniqueness. On the other hand, while MARS metrics (Table 4; columns 5-6) do not provide a clear-cut answer as to which classifier is individually strongest, they do bring forth valuable insight about the models' decision boundaries and possible synergies.

MARS ShineThrough (ST) and Occlusion (OCC) scores (Table 4; columns 5 and 6, respectively) and MARS charts (Figures 2-4) reveal that C_1 is uniquely adept at spotting one third (0.33) of the target class items, and, that while C_4 performs reasonably well on its own (Table 4; row 4), it could be used alongside C_1 to further optimize target-class item discovery. Occlusion scores further validate the combination of C_1 and C_4, as C_1 is the only classifier that has an Occlusion score > 0 (Figure 2), indicating that it has a unique target-class prediction error (@ i = 2, Table 2) that may be best handled by a secondary model (C_4 in this case, as it has the second highest ST score after C1).

While some classifier combinations may improve overall target-class discovery performance, the opposite is also possible. For example, Figure 2 shows that the combination of C_3 and C_4 produces MARS ShineThrough scores identical to those of C_4 alone, indicating that it is a weak combination, and should, therefore, be avoided. Thus, while traditional performance metrics gauge individual classifier capabilities by quantitively interpreting classifier-data interactions, MARS scores and charts examine classifier uniqueness and target-class discovery power by simultaneously interpreting both classifier-data and classifier-classifier interactions.

Note that the MARS evaluation mechanism was developed for a prototypical application of maximizing the volume of safety concerns found in online reviews, while constraining the close-reading verification effort required to determine if predicted positives are true positive. That is, the MARS method assists with elevating binary classifier yield: that is, increasing verified true positives per unit of effort reviewing predicted positives. The MARS evaluation mechanism is best suited to applications where the false positive cost is low, such as our prototypical application of discovering safety concerns in online reviews: a true positive (online review that contains a safety concern) is valuable, while a false positive (online review that does not contain a safety concern) has low cost, as each false positive wastes only a little reading effort, especially when there are few online reviews (predicted positives) shortlisted by the ML algorithm(s) for escalated attention by a human reviewer who is manually reviewing the predicted positive observations. For other applications – such as disease discovery – where the false positives, and false negatives, have differing trade-offs, the MARS evaluation method presented here may not be appropriate, and an inverted MARS evaluation method, aimed at maximizing true negatives, may be preferable.

Conclusions

In this paper, we presented the mathematical background and interpretation for two novel binary classification performance metrics – MARS ShineThrough and MARS Occlusion scores, whose software-level implementation, in the Python language, was recently described in Ref. 19. The formal definition of the MARS method, provided in this paper, will allow the research community to verify the correctness of the MARS method (through peer-review), accurately implement the MARS method in other programming languages (such as JavaScript, PHP, and R), and develop novel alternatives to, and enhancements to, the MARS method (such as visualizations that chart MARS metrics across multiple classifier cut-off thresholds instead of the single classifier cut-off threshold illustrated here). The stylized dataset and worked sample calculations provided in the Use cases section of this paper, above, is usable by the research community as a test case, to validate the correctness of each computational step of future software implementations. MARS metrics and MARS charts add yet another layer to the process of classifier assessment, providing crucial insight about each classifier's behavior relative to that of its peers. ShineThrough scores evaluate the comparative unique strengths of the classifier, by determining the proportion of total true positives that were exclusively found by the classifier. On the other hand, Occlusion scores measure the proportion of observations that were correctly labelled by the other classifiers but misclassified by the current classifier, i.e., the classifier's comparative unique weaknesses.

Naturally, the metrics synergize well with conventional measures, as the latter are constrained to the individual classifier's confusion matrix, while the former make use of the entire observation sample space, thus, evaluating classifier behavior from a previously unseen standpoint: the relative number of target class observations spotted or missed only (i.e., exclusively) by one classifier. This was demonstrated throughout the provided worked-out examples, which calculated ShineThrough and Occlusion scores for our stylized dataset (Table 2), and in Ref. 19 with a real dataset, albeit without the comprehensive mathematical explanation and examples presented in this paper. As a result, the MARS methodological framework adds a new classifier-comparison dimension – exclusive hits and misses – not expounded by conventional classifier evaluation methods.

Data availability

All data underlying the results are available as part of the article and no additional source data are required.

Software availability

Webapp: https://mars-classifier-evaluation.herokuapp.com

Source code available from: https://github.com/SoftwareImpacts/SIMPAC-2021-191

Archived source code at time of publication: https://doi.org/10.24433/CO.2485385.v1[20]

License: MIT

References

1. Mendez KM, Reinke SN, Broadhurst DI: **A comparative evaluation of the generalised predictive ability of eight machine learnng algorithms across ten clinical metabolomics data sets for binary classification.** *Metabolomics.* 2019; **15**: 150–150.
PubMed Abstract | Publisher Full Text
2. Hosenie Z, Lyon RJ, Stappers BW, *et al.*: **Comparing multiclass, binary, and hierarchical machine learning classification schemes for variae stars.** *Mon. Not. R. Astron. Soc.* 2019; **488**: 4858–4872.
Publisher Full Text
3. Makhtar M, Neagu DC, Ridley MJ: **Binary Classification Models Comparison: On the Similarity of Datasets and Confusion Matrix for Predictive Toxicology Applications.** *Inf. Technol. Bio- Med. Informatics.* 2011; 108–122.
Publisher Full Text
4. Mostafa FB, Hasan E: **Machine Learning Approaches for Binary Classification to Discover Liver Diseases using Clinical Data.** *MedRxiv.* 2021.
5. Narassiguin A, Bibimoune M, Elghazel H, *et al.*: **An extensive empirical comparison of ensemble learning methods for binary classification.** *Pattern Anal. Appl.* 2016; **19**: 1093–1128.
Publisher Full Text
6. Winkler M, Abrahams AS, Gruss R, *et al.*: ***TOY SAFETY SURVEILLANCE FROM ONLINE REVIEWS.*** *Decis. Support. Syst.* 2016; **90**: 23–32.
PubMed Abstract | Publisher Full Text
7. Abrahams AS, Fan W, Wang GA, *et al.*: **An Integrated Text Analytic Framework for Product Defect Discovery.** *Prod. Oper. Manag.* 2015; **24**: 975–990.
Publisher Full Text
8. Goldberg DM, Khan S, Zaman N, *et al.*: **Text Mining Approaches for Postmarket Food Safety Surveillance Using Online Media.** *Risk Anal.* 2020.
PubMed Abstract | Publisher Full Text
9. Adams DZ, Gruss R, Abrahams AS: **Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews.** *Int. J. Med. Inform.* 2017; **100**: 108–120.
PubMed Abstract | Publisher Full Text
10. DMW Powers: 2020.
11. Altman DG, Bland JM: **Diagnostic tests. 1: Sensitivity and specificity.** *BMJ.* 1994; **308**: 1552–1552.
Publisher Full Text
12. Chinchor N: **MUC-4 Evaluation Metrics**. *Proc. 4th Conf. Messag. Underst.* Association for Computational Linguistics; 1992; pages 22–29.
13. Sasaki Y: 2007.
14. Sokolova M, Japkowicz N, Szpakowicz S: **Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation BT - AI**. Sattar A, Kang B, editors. *Advances in Artificial Intelligence.* Springer; 2006; pages 1015–1021.
15. Van Rijsbergen C: **Information retrieval: theory and practice.** *Proc. Jt. IBM/University.* 1979; pages 1–14.
16. Hanley JA, Mcneil BJ: **The meaning and use of the area under a receiver operating characteristic (ROC) curve.** *Radiology.* 1982; **143**: 29–36.
PubMed Abstract | Publisher Full Text
17. Bradley AP: **The use of the area under the ROC curve in the evaluation of machine learning algorithms.** *Pattern Recogn.* 1997; **30**: 1145–1159.
Publisher Full Text
18. Saito T, Rehmsmeier M: **The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets.** *PLoS One.* 2015; **10**: e0118432.
PubMed Abstract | Publisher Full Text
19. Mali N, Restrepo F, Abrahams A, *et al.*: **Implementation of mars metrics and Mars charts for evaluating classifier exclusivity: The comparative uniqueness of binary classifier predictions.** *Software Impacts.* 2022; **12**: 100259.
Publisher Full Text
20. Mali N, Restrepo F, Abrahams A: **Implementation of MARS metrics and MARS charts for evaluating classifier exclusivity: the comparative uniqueness of binary classifier predictions [Source Code].** 2021.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 03 August 2022

https://doi.org/10.5256/f1000research.135201.r142931

✔ **Samir Chatterjee**
School of Information and Technology Management, Claremont Graduate University, Claremont, USA

I am happy with the changes.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: ML in Healthcare

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 18 July 2022

https://doi.org/10.5256/f1000research.135201.r142932

✔ **Timothy A. Warner**
Department of Geology and Geography, West Virginia University, Morgantown, WV, USA

Many thanks for the careful revisions and particularly for the detailed response to my suggestions. The revised paper makes a valuable contribution.

Some very minor, final suggestions are listed below:

1. The MARS acronym in the abstract (p1) and the text (p3) seem to be different: the latter seems to be missing “Specificity” at the end.

2. In the abstract, I suggest adding “may” in front of “share nearly identical performance” (because classifier performance is sometimes not nearly identical).

3. In the second sentence of the last paragraph on p3, I suggest adding “Ref.” or something similar to make the sentence easier to read. (E.g., so that it would read, “Recently, Ref. 1 evaluated…”).

4. P6, in the paragraph below the two numbered paragraphs: I suggest replacing “calculations are identical” with “calculations are similar”.

5. Equation 2. Shouldn’t the variable *i* following *Z* be a subscript?

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Remote sensing

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 16 June 2022

https://doi.org/10.5256/f1000research.122189.r136525

?

Samir Chatterjee
School of Information and Technology Management, Claremont Graduate University, Claremont, USA

The paper proposes two new binary classifier metrics in addition to existing traditional metrics such as accuracy, precision, recall, F-score. Two classifiers of equal accuracy may each have the unique ability to identify distinct observations from the target class.

MARS ShineThrough and MARS Occlusion scores are mathematically presented.

On page 7, where MARS occlusion scores is first defined, it should read "total number of expected false negatives (EFNcj)."

In ML, combining algorithms is not too common. While ensemble methods use a similar core algorithms but creates different variations of the classifier (many trees in a Random Forest

implementation), it may not be practically feasible to combine classifiers that are inherently built using different algorithms (Logistic regression, SVM, KNN). In those cases, how will this technique apply?

The visual bubble graphs are also little hard to understand.

When comparing between different classifiers, while ST and OCC may tell us unique distinguishing capability of the classifiers, it is also important to have a discussion of the consequences of the TP and FN especially when it might be related to disease predictions. What are the trade-offs?

Is the rationale for developing the new method (or application) clearly explained?
Partly

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: ML in Healthcare

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 17 Jun 2022

Peter Ractham, Thammasat University, Bangkok, Thailand

Manuscript Number: 110567

Dear Dr. Chatterjee,

Thank you for your helpful comments and suggestions regarding our manuscript, *Formal definition of the MARS method for quantifying the unique target class discoveries of selected machine classifiers*. We have edited the manuscript with the corrections and clarifications

prompted by your suggestions.

Below this letter, we have addressed the comments in a point-by-point manner. Our response, **bolded**, includes a brief explanation behind the initial reasoning and how and where the paper was modified.

Thank you for your time and consideration.

Sincerely,

Felipe Restrepo
Namrata Mali
Alan Abrahams
Peter Ractham

Comments to the Authors:

1. On page 7, where MARS occlusion scores is first defined, it should read "total number of expected false negatives (EFN_{Cj})

Thank you for noticing this. We have made this fix.

1. In ML, combining algorithms is not too common... it may not be practically feasible to combine classifiers that are inherently built using different algorithms (Logistic regression, SVM, KNN). In those cases, how will this technique apply?

We have clarified in the manuscript that we are not proposing that the algorithm procedures be combined, but rather that the predicted positive observations of the algorithms be combined. Specifically, the intersection of the sets of predicted positives of the algorithms is taken, which is a straightforward operation. The combination of predicted positives, from two algorithms that each suggest distinct true positives, allows the data scientist to efficiently boost the number of true positives while constraining total positive predictions, which is highly desirable for applications like our prototypical application: maximizing the volume of safety concerns found in online reviews, while minimizing the close-reading verification effort required to determine if a predicted positive is a true positive.

1. The visual bubble graphs are also little hard to understand.

Thank you for your suggestion. To assist readers who find bubble charts difficult to interpret, we have supplemented the visual bubble graphs with MARS bar charts, which show the total distinct true positives, for each algorithm, sorted from algorithm with most distinct true positives down to the algorithm with the least distinct true positives.

1. It is also important to have a discussion of the consequences of the TP and FN especially when it might be related to disease predictions. What are the trade-offs?

Thank you for pointing out the importance of adding this discussion. We have added the following: "The MARS evaluation mechanism was developed for a prototypical application of maximizing the volume of safety concerns found in online reviews, while constraining the close-reading verification effort required to determine if predicted positives are true positive. That is, the MARS method assists with elevating binary classifier yield: that is, increasing verified true positives per unit of effort

reviewing predicted positives. The MARS evaluation mechanism is best suited to applications where the false positive cost is low, such as our prototypical application of discovering safety concerns in online reviews: a true positive (online review that contains a safety concern) is valuable, while a false positive (online review that does not contain a safety concern) has low cost, as each false positives wastes only a little reading effort, especially when there are few online reviews (predicted positives) shortlisted by the ML algorithm(s) for escalated attention by a human reviewer who is manually reviewing the predicted positive observations. For other applications – such as disease discovery – where the false positives, and false negatives, have differing trade-offs, the MARS evaluation method presented here may not be appropriate, and an inverted MARS evaluation method, aimed at maximizing true negatives, may be preferable.

Thank you again for your helpful observations and suggestions, which have helped us improve the clarity of the manuscript.

Competing Interests: No competing interests were disclosed.

Author Response 17 Jun 2022

Peter Ractham, Thammasat University, Bangkok, Thailand

Dear Reviewer,

It seems that we'll have to wait for the update version of the manuscript from F1000 before we can submit the revised manuscript suggested by you. We'll update it as soon as possible. Thank you.

Competing Interests: No competing interests were disclosed.

Reviewer Report 17 May 2022

https://doi.org/10.5256/f1000research.122189.r136529

?

Timothy A. Warner
Department of Geology and Geography, West Virginia University, Morgantown, WV, USA

This paper describes statistics that summarize similarity of the labelling of unknown samples by different classifiers. The method has two levels: (1) individual classifiers vs the rest and (2) groups of two classifiers vs the rest. The method and the visualization of the results were described in an earlier paper; this paper provides a more thorough definition of the statistics and a worked,

hypothetical example.

The fact this paper has the express aim of providing clarification to an earlier paper may constrain how open the authors are to modifying their approach.

Major suggestions

1. The two measures of shine through and occlusion are described numerous times as respectively measures of exclusive true positives and exclusive false negatives. However, the definition of exclusive seems to differ in the two cases. For shine through, exclusive means true positives (TP) for *only* the classifier of interest, and no other classifier (i.e. FN for all other classifiers). For occlusion exclusive means false negatives (FN) for the classifier of interest, but a TP for *any* (i.e., at least one) other classifier. Thus for occlusion the word "exclusive" has a more relaxed meaning than for shine through. If the meaning of exclusive for occlusion were the same as for shine through, then the definition of occlusion would be a FN for the classifier of interest, and no other classifier.
2. The paper focuses on "exclusive" TP and FP. However, as Figure 1 from the paper shows, there are two other types of classification outcomes: false positives (FP) and true negatives (TN). A key aspect of classifier behavior is the trade-off between FN and FP. Is there any reason for not developing similar metrics for FP and TN, and thus providing a comprehensive, instead of partial, view of the differences between classifiers?
3. The paper makes a valuable contribution in providing statistical measures that compare decision boundaries between classifiers. However, I think it is potentially confusing to suggest that the MARS statistics are "alternative classification performance metrics" that overcome "limitations" of "traditional performance metrics." I think that is a bit like saying the problem with the mean as a measure of central tendency is that it doesn't measure autocorrelation. The MARS statistics don't seem to be performance metrics, in the sense of quantifying accuracy. I think being clear about the purpose and role of MARS statistics is important in helping readers understand what information the MARS can offer.
4. Based on the above, I suggest removing the extensive discussion of Kappa, F-score, MCC, AUC, ROC, PR curves, and class imbalance, which seems to be a distraction. Furthermore, it seems to me class imbalance will affect MARS metrics as much as any other statistic, so I don't follow the argument that MARS represents a method to address this limitation in conventional accuracy statistics. I think all you need to say is that two classifiers can have the same summary accuracy statistics (such as overall accuracy, precision and recall), but have different decision boundaries. MARS helps one explore those differences. Similarly, in the discussion, I suggest emphasizing that the power of the MARS measures is not in clarifying the accuracy of the various classifiers, but rather in highlighting differences in their decision boundaries.
5. For the MARS charts, I suggest following the example of a covariance matrix (where the variance is on the diagonal, and covariance on off-diagonal positions), and place the single classifier values on the diagonal, and the classifier combinations on the off-diagonal positions. (However, I suggest keeping the different colors, which I found useful.) I think the use of the diagonal and off-diagonal in this way is conceptually clearer, and also has the benefit that you don't have the problem of which one to prioritize when the two circles have the same diameter.
6. I don't understand Table 4. The numbers in table 4 for C1,4 seem to be a duplicate of C1 in Table 2, and C2,3 a duplicate of C2. Crucially, I can't relate it to the calculation of ETPsub(C1,4), nor does it seem to agree with table 5.

7. In table 5, I don't understand how the overall accuracy, precision and recall for the combination of classifiers (C1,4 and C3,4) are calculated. For example, I don't see what objective rule combining C1 and C4 could result in class labels that indicate 100% accuracy for these samples.

Minor suggestions

1. Perhaps explain the MARS acronym? I don't understand the reference to sensitivity and specificity in the context of shine through and occlusion (especially since specificity is the TN as a proportion of the reference Negative class; the current MARS statistics do not seem to include an "exclusive" TN measure).
2. I suggest defining MARS acronym in the main body – currently it seems to be only defined in the abstract. (Unless I missed it. If so, sorry.)
3. Second sentence under the heading "Related work" – I suggest that in defining the recall, add the word "actual" prior to "positives", so that the definition becomes "the overall proportion of *actual* positives that were correctly labelled as such."
4. "Accuracy" – defined on p5. I suggest using the term "overall accuracy" rather than just "accuracy" to differentiate this from the generic concept of accuracy.
5. I think the equations would be easier to follow if you moved the reference to Table 1 to the start of the section with the equations.
6. I suggest not using the same symbol for more than one purpose. For example, the constant Z-sub-i has different definitions in 2.1 and 4.1. (When I first read the paper, I incorrectly used the 2.1 definition when I was working through the occlusion example. Using a different letter for the 2 constants would avoid this problem.)
7. Similarly, in eqn 2, it was a bit confusing to me that subscript j on the left side of the equation could (in fact, has to) simultaneously represent a different value on the right hand side of the equation. Using a different symbol on the left side will obviate this confusion.
8. P8 – Second-last paragraph "To calculate individual Shine through scores"....I suggest referring to Table 3 here to clarify how ETP is calculated.
9. P9 –Has variable k been defined? Is it perhaps Z-sub-i?
10. P9. The example of occlusion for C1, @i=1. In the first worked example, is the first 0 and 1 (y11 and t1) switched? I.e, it seems to me that this should read "max (1 x 0, 0 x 0, 0 x 0) x 0 = 0"?
11. Table 3. The value for Z-sub-i for observation 1 is listed in the table as 1. Should it not be 0?
12. End of first paragraph below Table 3. The reference to "Tables 1 and 3" – shouldn't this be to "Tables 1 and 4"?
13. MARS charts – the discussion and captions simplifies the MARS metrics as "counts" – but they are actually defined as proportions. I think adding a legend that indicates how circle size relates to proportions would be useful.

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Remote sensing

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 16 Jun 2022

Peter Ractham, Thammasat University, Bangkok, Thailand

Manuscript Number: 110567

Dear Dr. Warner,

Thank you for reviewing our manuscript, *Formal definition of the MARS method for quantifying the unique target class discoveries of selected machine classifiers*. We greatly appreciate the careful review and valuable feedback you have provided. We have incorporated the suggested changes to the paper and software artifact.

Below this letter, we have addressed the comments in a point-by-point manner. Our response, **bolded**, includes a brief explanation behind the initial reasoning and how and where the paper was modified.

Thank you for your time and consideration.

Sincerely,

Felipe Restrepo
Namrata Mali
Alan Abrahams
Peter Ractham

Comments to the Authors:

1. The two measures of shine through and occlusion are described numerous times as respectively measures of exclusive true positives and exclusive false negatives.

However, the definition of exclusive seems to differ in the two cases. For shine through, exclusive means true positives (TP) for *only* the classifier of interest, and no other classifier (i.e. FN for all other classifiers). For occlusion exclusive means false negatives (FN) for the classifier of interest, but a TP for *any* (i.e., at least one) other classifier. Thus for occlusion the word "exclusive" has a more relaxed meaning than for shine through. If the meaning of exclusive for occlusion were the same as for shine through, then the definition of occlusion would be a FN for the classifier of interest, and no other classifier.

Thank you for pointing this out. We have modified the definition of occlusion so that the definition of exclusive true positives now directly parallels that of exclusive false negatives. The updated mathematical formulation is now:

Formula 1:
https://f1000researchdata.s3.amazonaws.com/linked/432866.formula1.png

Formula 2:
https://f1000researchdata.s3.amazonaws.com/linked/432867.formula2.png

Following this formulation, Occlusion scores now represent the proportion of the classifier's unique misses, relative to discovered true positives. A high Occlusion score may suggest that the classifier has a rather specific struggle within the classification task, which the remaining models are capable of handling.
Within the manuscript, we updated the MARS Occlusion Scores section (p. 5) with the revised mathematical formulation, the MARS Occlusion Score Metric: Example Computation section (p. 8) with the revised mathematical formulation, and Table 4 (p.10) Occlusion scores. Figure 3 (MARS Occlusion Chart) in the MARS charts section (p. 9) was also updated to reflect the updated definition.

1. The paper focuses on "exclusive" TP and FP. However, as Figure 1 from the paper shows, there are two other types of classification outcomes: false positives (FP) and true negatives (TN). A key aspect of classifier behavior is the trade-off between FN and FP. Is there any reason for not developing similar metrics for FP and TN, and thus providing a comprehensive, instead of partial, view of the differences between classifiers?

Thank you for pointing this out. Given that MARS metrics were designed as a tool to optimize big-data, machine-driven discovery efforts in which the value of TPs and the cost of FNs is high, - e.g., flagging potentially hazardous products via online reviews – we focused on defining MARS through TPs and FNs. The approach can easily be adapted to applications in which the value and cost of TNs/FPs is high.
Within the manuscript, we have clarified that the approach is easily replicated with TNs and FPs, but we omit detailed calculations (as they would be nearly identical to those of TPs and FNs) for brevity (Methods section, p. 4).

1. The paper makes a valuable contribution in providing statistical measures that compare decision boundaries between classifiers. However, I think it is potentially confusing to suggest that the MARS statistics are "alternative classification performance metrics" that overcome "limitations" of "traditional performance metrics." I think that is a bit like saying the problem with the mean as a measure of central tendency is that it doesn't measure autocorrelation. The MARS statistics don't

seem to be performance metrics, in the sense of quantifying accuracy. I think being clear about the purpose and role of MARS statistics is important in helping readers understand what information the MARS can offer.

Thank you for your suggestion. We intended for MARS metrics to be used alongside traditional metrics as a tool to optimize big-data, machine-driven discovery efforts. MARS scores were designed to complement traditional ones (e.g., accuracy, recall, precision) in high-volume data applications, where models are likely to have similar conventional summary statistics, even if their decision boundaries are fundamentally different. In these cases, the depth of traditional metric analysis may be significantly limited, as results would simply suggest that all models employed worked well on the data, providing no differentiating power within the set of classifiers. Whereas using MARS metrics, which are far more likely to detect differences in classifier behavior, alongside traditional metrics, would allow for a more complete analysis to be made about the model's overall (individual and comparative) performance.
Within the manuscript, we have modified the Introduction section (p. 2 – 3) so that the purpose and role of MARS statistics is clear for readers.

1. Based on the above, I suggest removing the extensive discussion of Kappa, F-score, MCC, AUC, ROC, PR curves, and class imbalance, which seems to be a distraction. Furthermore, it seems to me class imbalance will affect MARS metrics as much as any other statistic, so I don't follow the argument that MARS represents a method to address this limitation in conventional accuracy statistics. I think all you need to say is that two classifiers can have the same summary accuracy statistics (such as overall accuracy, precision and recall), but have different decision boundaries. MARS helps one explore those differences. Similarly, in the discussion, I suggest emphasizing that the power of the MARS measures is not in clarifying the accuracy of the various classifiers, but rather in highlighting differences in their decision boundaries.

Thank you for your suggestion. The class imbalance discussion was meant to highlight the vulnerabilities of conventional metrics and bring forth the need for novel metrics capable of examining classifiers from a different standpoint. We did not intend to imply that MARS represents a method to address this limitation. Rather, we intended to express that conventional metrics would greatly benefit from the use of MARS alongside, as doing so would allow for a more objective and in-depth analysis of the model's behavior – which we have now made clear in the Introduction section, rendering the class imbalance discussion unnecessary.
Within the manuscript, we have removed the class imbalance discussion in the Introduction section (p. 3) and significantly shortened the Related Works section by doing the same (p. 3-4). Within Related Works (p. 3-4), we also reduced the discussion pertaining to PR and ROC curves. The Discussion section (p. 10) was reworked to better emphasize the power of MARS metrics in spotting differences between classifier behavior and optimizing model combinations.

1. For the MARS charts, I suggest following the example of a covariance matrix (where the variance is on the diagonal, and covariance on off-diagonal positions), and place the single classifier values on the diagonal, and the classifier combinations on the off-diagonal positions. (However, I suggest keeping the different colors, which I found useful.) I think the use of the diagonal and off-diagonal in this way is conceptually clearer, and also has the benefit that you don't have the problem of which one to prioritize when the two circles have the same diameter.

Thank you for your suggestion. We have implemented the suggested changes to the MARS charts and updated Figures 2 and 3 (Mars charts section, p. 9-10).

1. I don't understand Table 4. The numbers in table 4 for C1,4 seem to be a duplicate of C1 in Table 2, and C2,3 a duplicate of C2. Crucially, I can't relate it to the calculation of ETPsub(C1,4), nor does it seem to agree with table 5.

Thank you for pointing this out. We have removed Table 4 from the manuscript. Upon review, it does not align with the reworked Discussion and Introduction sections, as combined MARS metrics are meant to facilitate the discovery of classifiers with complementary decision boundaries; they are not designed for traditional classifier ensemble creation, as Table 4 suggested.

1. In table 5, I don't understand how the overall accuracy, precision and recall for the combination of classifiers (C1,4 and C3,4) are calculated. For example, I don't see what objective rule combining C1 and C4 could result in class labels that indicate 100% accuracy for these samples.

Thank you for pointing this out. As per the previous suggestion, we re-evaluated Table 4 and determined it was best eliminated. Consequently, C1,4 and C3,4 have been removed from Table 5.

1. Perhaps explain the MARS acronym? I don't understand the reference to sensitivity and specificity in the context of shine through and occlusion (especially since specificity is the TN as a proportion of the reference Negative class; the current MARS statistics do not seem to include an "exclusive" TN measure).

Thank you for your suggestion. Since sensitivity examines the model's target-class detection capabilities with respect to the complete TP sample space, we referenced MARS as a tool to determine relative sensitivity, as it examines the model's target-class detection capabilities with respect to competing classifiers, rather than the "ground truth". We followed the same logic when referring to relative specificity. However, since we only briefly mention the ease of adaptability to TNs and FPs (refer to revision 2.), we have removed the specificity reference (Abstract, p. 1 & Introduction, last paragraph).

1. I suggest defining MARS acronym in the main body – currently it seems to be only defined in the abstract. (Unless I missed it. If so, sorry.)

Thank you for your suggestion. We had previously defined the MARS acronym in the main body (Introduction, last paragraph), and have now added an additional definition when first mentioned (Introduction, p. 2 ¶ 3).

1. Second sentence under the heading "Related work" – I suggest that in defining the recall, add the word "actual" prior to "positives", so that the definition becomes "the overall proportion of *actual* positives that were correctly labelled as such."

Thank you for your suggestion. We have added "actual" prior to positives (Related work, p. 3, ¶ 1).

1. "Accuracy" – defined on p5. I suggest using the term "overall accuracy" rather than just "accuracy" to differentiate this from the generic concept of accuracy.

Thank you for your suggestion. We have added "overall" prior to accuracy (Related Work, p. 4, ¶ 5).

1. I think the equations would be easier to follow if you moved the reference to Table 1 to the start of the section with the equations.

Thank you for your suggestion. We have moved Table 1 to the start of the Methods

section (p. 5).

1. I suggest not using the same symbol for more than one purpose. For example, the constant Z-sub-i has different definitions in 2.1 and 4.1. (When I first read the paper, I incorrectly used the 2.1 definition when I was working through the occlusion example. Using a different letter for the 2 constants would avoid this problem.)

Thank you for your suggestion. We have changed the constant Z-sub-i to R-sub-i for the Occlusion metric.

1. Similarly, in eqn 2, it was a bit confusing to me that subscript j on the left side of the equation could (in fact, has to) simultaneously represent a different value on the right hand side of the equation. Using a different symbol on the left side will obviate this confusion.

Thank you for your suggestion. We have added C_w which represents the classifier of interest and resolves the confusion. Within the manuscript, we updated equations 1 – 5 (Methods section, p. 5 – 6) to reflect this change. The new term was also added to Table 1 (p. 5).

1. P8 – Second-last paragraph "To calculate individual Shine through scores"....I suggest referring to Table 3 here to clarify how ETP is calculated.

Thank you for pointing this out. We have referenced Table 3 for the worked-out ETP calculation example (MARS ShineThrough score metric: example computation, p. 7).

1. P9 –Has variable k been defined? Is it perhaps Z-sub-i?

Thank you for pointing this out. Yes, it was meant to be Z-sub-i (now R-sub-i for Occlusion scores). Within the manuscript we have substituted all mentions of k for R-sub-i (MARS occlusion score metric: example computation, p. 8 -9)

1. P9. The example of occlusion for C1, @i=1. In the first worked example, is the first 0 and 1 (y11 and t1) switched? I.e, it seems to me that this should read "max (1 x 0, 0 x 0, 0 x 0) x 0 = 0"?

Thank you for pointing this out. Yes, it should be '1 x 0' instead of '0 x 1'. We have corrected the order within the manuscript (MARS occlusion score metric: example computation, p. 8).

1. Table 3. The value for Z-sub-i for observation 1 is listed in the table as 1. Should it not be 0?

Thank you for pointing this out. Yes, it should be zero. We have made the change in Table 3 (MARS ShineThrough score metric: example computation, p. 7).

1. End of first paragraph below Table 3. The reference to "Tables 1 and 3" – shouldn't this be to "Tables 1 and 4"?

Thank you for pointing this out. The initial reference should have been Tables 1 and 4. However, since Table 4 was removed, it has been updated to 'Table 1' (MARS occlusion score metric: example computation, p. 9).

1. MARS charts – the discussion and captions simplifies the MARS metrics as "counts" – but they are actually defined as proportions. I think adding a legend that indicates how circle size relates to proportions would be useful.

Thank you for your suggestion. We have added an additional line to the captions of Figures 2 and 3 (Mars charts section, p. 8 – 9) explaining that bubble size is proportional to ShineThrough/Occlusion score: the larger the bubble, the higher the classifier(s) ShineThrough score.

Competing Interests: No competing interests were disclosed.