UDC 519.6; 662.2

L.M. DESCHAINE, P. NORDIN, JA.D. PINTER

A COMPUTATIONAL GEOMETRIC / INFORMATION THEORETIC METHOD TO INVERT PHYSICS-BASED MEC MODELS ATTRIBUTES FOR MEC DISCRIMINATION

Анотація. *Наявність залишкових підповерхневих боєприпасів і вибухових речовин (БВР) є серйозною проблемою в усьому світі. Дискримінація БВР від не БВР-елементів дозволяє спрямовувати ресурси на пом'якшення ризиків. Збір фізичних даних і інвертування, моделей, що фізично визначаються, проводяться з наміром використовувати інвертовані модельні параметри як базис для дискримінації БВР. Однак дискримінація БВР через модельну інверсію стикається зі значними труднощами в середовищах з шумами, а також при невизначеному місцезнаходженні сенсорів. Наш обчислювально-геометричний підхід демонструє можливість отримувати безліч інформаційних атрибутів, корисних для БВР-дискримінації, включаючи інформаційний зміст інвертованої моделі разом з цінною додатковою інформацією, недоступною при використанні інверсного підходу.*
Ключові слова: *боєприпаси і вибухові речовини, метод обчислювальної геометрії, техніка інверсії фізичної моделі.*

Аннотация. *Наличие остаточных подповерхностных боеприпасов и взрывчатых веществ (БВВ) является серьезной проблемой во всем мире. Дискриминация БВВ от не БВВ-элементов позволяет направлять ресурсы на смягчение рисков. Сбор физических данных и инвертирование физически определяемых моделей производятся с намерением использовать инвертированные модельные параметры в качестве базиса для дискриминации БВВ. Однако дискриминация БВВ через модельную инверсию сталкивается со значительными трудностями в средах с шумами, а также при неопределенном местоположении сенсоров. Наш вычислительно-геометрический подход демонстрирует возможность получать множество информационных атрибутов, полезных для БВВ-дискриминации, включая информационное содержание инвертированной модели вместе с ценной дополнительной информацией, недоступной при использовании инверсного подхода.*
Ключевые слова: *боеприпасы и взрывчатые вещества, метод вычислительной геометрии, техника инверсии физической модели.*

Abstract. *The presence of subsurface munitions and explosives of concern (MEC) is a significant issue worldwide. Discrimination of MEC from non-MEC items enables resources be focused on mitigating risk. Geophysical data is collected and physically-based models inverted with the intent that the inverted model parameters form the basis for MEC discrimination. However, MEC discrimination via model inversion has significant difficulties in noisy environments and with uncertain sensor location. Our computational geometric approach is demonstrated to produce an information-rich set of attributes useful for MEC discrimination including the inverted model information content along with valuable additional information not obtainable using the inversion approach.*
Keywords: *munitions and explosives of concern, computational geometric method, physics model inversion technique.*

1. Introduction

Solving MEC discrimination decision problems requires an in-depth understanding of the underlying science of geophysics. Our overall goal is to demonstrate the enhanced accuracy and performance possible from using machine learning modeling to fuse the information content obtained from MEC feature attributes derived from both data-driven models (using computational geometry) and physics-based models. We describe the techniques, and how the machine-learning independent information-theoretic approach can be used to assess the contribution from each feature source (computational geometry or the fitted physics models) in MEC discrimination challenge. The physics-based governing equations provide the relevant scientific problem space of

MEC item responses to geophysical interrogation. Computational geometry provides attributes for MEC and non-MEC (i.e. clutter, shrapnel). Hence, a key objective of this work is to merge and extend the techniques, effectively fusing both a priori physics-based and automatic modeling-based components to extend the maximum total discrimination/classification accuracy beyond that achievable by either method used independently. A related and equally important objective is to quantify the relative value of each component of the information sources in relationship to accuracy.

2. Overview of MEC Discrimination

MEC discrimination presents one of the toughest and most challenging problems in the genre of subsurface identification tasks. A MEC item can, for instance, be unexploded ordnance of various sizes and be buried below ground (fig. 1). MEC can retain their ability to detonate; they pose a continuing risk. The United States Department of Defense (DOD) has invested heavily in basic research and development to address this challenge, but because typically MEC targets are small and surrounded by clutter (e.g., shrapnel or non-MEC items), accurate and reliable discrimination has been a challenge. Hence, while progress is being made, safe, efficient and cost-effective solutions have so far proven elusive.

Fig. 1. Typical MEC and non-MEC items. (Image: US Army Environmental Command: Standardized Target Specifications: Technology Demonstration Sites)

Initially, MEC discrimination research focused on two primary approaches to evaluate a Target of Interest (TOI): the first, a physics-based approach [1], relied on mathematical models whereby model parameters were fitted to field data by solving the inverse modeling problem. A second approach, which used machine-learning modeling and multidisciplinary computational geometry insights to derive features from the field data, clearly outperformed the other methods in use at that time to discriminate MEC from non-MEC [2]. Both approaches are described below.

2.1. Inverse (Fitted) Physics-Based Models

This section explains the inverse physics-based modeling approach for discriminating MEC items using electromagnetic (EMI)-based and magnetic (MAG) instruments.

One method to investigate the presence of MEC items is by conducting non-destructive geophysical surveys. This approach has value only if the resulting information is useable for locating anomalies and discriminating between MEC and non-MEC items. Since the MEC objects are not observable (being primarily below ground), the location, depth, and orientation of the MEC item are unknown. These model parameters are solved for by inverse modeling and are used to assess whether a TOI is a MEC item or not.

EMI uses induction theory and leverages the hypothesis that the distributions of the eigenvalues of magnetic polarizability provide an understandable basis for MEC versus non-MEC discrimination. This hypothesis is based on the observation that a MEC item can be approximated by an axisymetric cylindrical (as illustrated on Fig. 1) and, therefore, has only two unique eigenvalues, one that represents the length of the object and the other two that represent the axial symmetry. Irregular objects (e.g., clutter), however, exhibit three distinct eigenvalues (that is, different responses in three orthogonal directions). The model of the signal $S(t)$ that is generated by the EMI equipment is:

$$S(t)=\frac{\partial}{\partial t}Tr\{\mathbf{TR}\cdot\mathbf{B}(t)\};\quad B(t)=\begin{bmatrix} b11(t) & \cdots & b13(t) \\ \vdots & \ddots & \vdots \\ b31(t) & \cdots & b33(t) \end{bmatrix}. \tag{1}$$

Where t is time, Tr is the sum of the diagonal elements of a matrix (trace), TR is the transmit/receive matrix, and $B(t)$ is a symmetric-effective polarizability matrix. $S(t)$ is computed from the convolution of the magnetic polarizability with the transmit waveform. The best-fit eigenvalues (β1, β2, β3) correspond to the responses induced when the primary field is aligned with the principal axes of the object. A magnetic (MAG) survey response is described by a simple dipole model. A tool that provides the best fit estimate for both EMI and MAG data (UX-Analyze) has been developed by ESTCP to facilitate these calculations [4]. Fig. 2 illustrates the results of an inverse model fit for an anomaly investigated using both the EMI and MAG geophysical techniques.

This approach provides fitted model parameters that are listed under the "fit results" output summary. There are seven EMI-fitted model parameters, which are then used as inputs for machine learning modeling: these are the depth of the object (Depth), its size (Size), the eigenvalues (β1, β2, β3), the Coh and the best-fit value (chi2). Inverse physics modeling for the MAG sensor provides as outputs depth, size, declination, inclination, solid angle, and the magnetic moment. MEC discrimination insight is gained from data collected later in the decay curve which captures the anomaly metal thickness. The core concept regarding the EMI inverse model technique is that the polarizability will have one large (β1) and two small (β2, β3) and equivalent values to describe the conical MEC-shaped item. MAG relies on the shape and amplitude aspects. Hence, both shape (cylindrical versus fragments) and metal thickness (casings versus sheet metal) are also useful MEC discrimination information.

While theoretically sound, significant practical challenges to this method include the need to overcome data collection positioning error (requires resolution on the centimeter scale); and signal-to-noise ratio (S/N) must be very high, on the order of 100, and non-uniqueness of the eigenvalue solutions. The inverse model parameters used in this work were developed by [5].

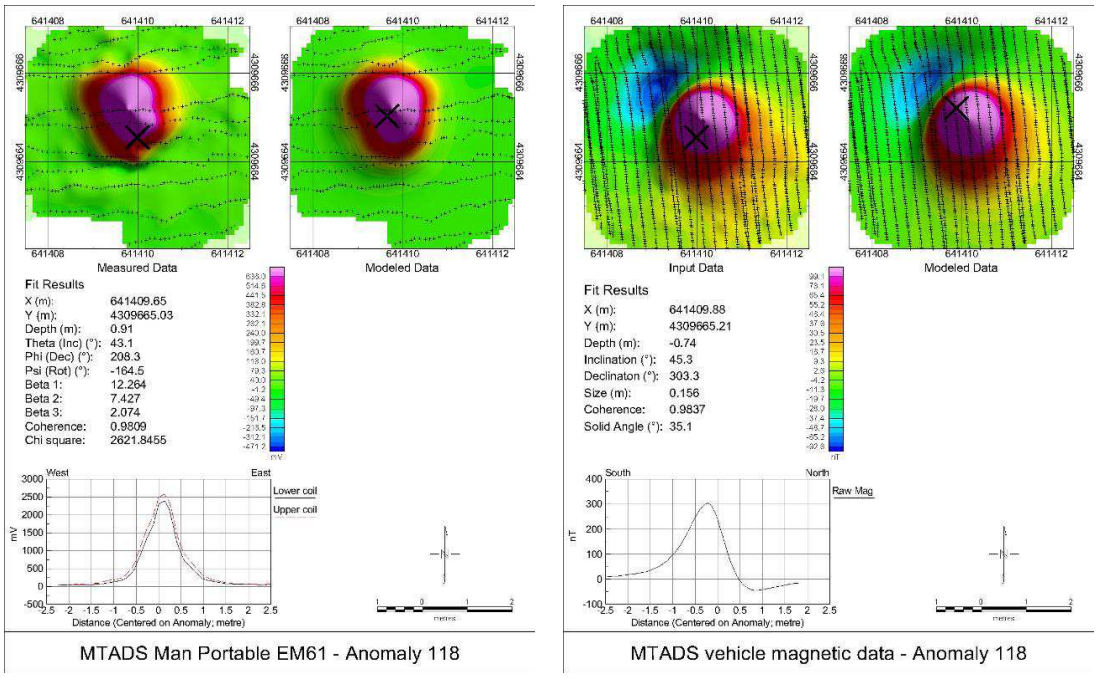

Fig. 2. Inverse modeling analysis of EMI (left) and MAG (right) for one anomaly using UX-Analyze. (From fig. 2–7 in [4])

2.2. Computational Geometric Model

We first developed and tested the multi-disciplinary, machine-learning approach using computational geometric modeling techniques in the fall of 2001 on publicly available information and data sets for a MEC (then called "UXO" for unexploded ordnance) discrimination from a "prove-out" site known as the Jefferson Proving Ground – Phase IV. The approach performed far better than any technique used at that time [2]. The data used were collected by others using a Protem-47, time domain geophysical unit that provided 20 time gates of signal S(t) information. The compiling genetic programming system (CGPS), a machine-learning technique developed by Nordin [6], was used as the classification algorithm (we later coined the phrase "linear genetic programming" [LGP] to differentiate it from other genetic programming algorithms). The results of this study are summarized in Deschaine [2] and are shown in fig. 3.

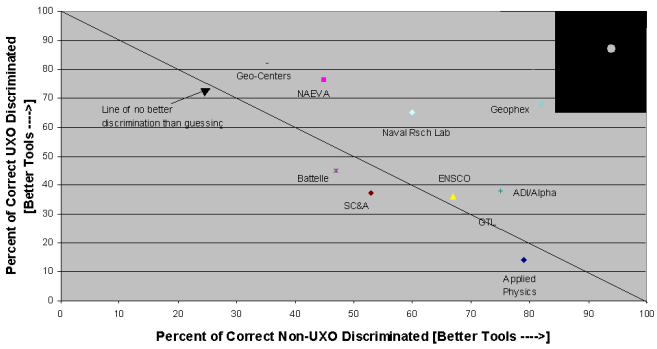

Fig. 3. MEC discrimination solution compared to results from the JPG Phase IV UXO (MEC) Discrimination Project

Fig. 3 shows the performance of the published results from 10 analyses conducted by vendors who provided MEC discrimination services as part of the JPG Phase IV project. The horizontal axis shows the performance of each method in correctly identifying anomalies that did not contain buried MEC; whereas the vertical axis shows the performance of each method in correctly identifying anomalies that did contain buried MEC. The angled line in the figure represents what could be expected from random guessing.

The difficulties of modeling these data are evident: most methods performed little better than random guessing would. Notwithstanding this limitation, the machine-learning based computational geometric approach using the CGPS algorithm still provided the best-known approach at the time for correctly identifying MEC and for correctly rejecting non-MEC using various data set configurations on blind data [2]. The dashed line from the NAVEA solution in Fig. 3 indicates that the data set for the machine-learning algorithm was used. Note that the data we used was from a well conducted study, yet the analysis method used by others only produced results slightly above average. (We selected this data because of its computational geometric value.) Note that we intentionally did not use the data set labeled Geophex, even though it had the best performance of the group as analyzed by others, because we concluded that the NAVEA data had more information for high accuracy MEC discrimination – the team doing the original analysis just were not able to exploit it. The gray dot in the upper right-hand corner of the figure shows the CGPS solution on unseen data. Because the number of data points was small, we used a re-sampling technique to estimate the 95% confidence interval on this solution; the black rectangle in Fig. 3 shows that interval. CGPS – combined with computational geometric approach – produced by far the most accurate discrimination results.

Since the initial UXO/MEC discrimination success in 2001, we have been assessing the challenge of quickly finding targets of interest and then extracting a small, focused set of MEC/non-MEC relevant discrimination features for input to machine-learning algorithms and production-size data sets. The initial approach we tested was to use genetic programming for automated feature extraction, but it was unsuccessful in practice. The approach we found that is robust, practical, flexible, and effective is a multi-disciplinary formulation of computational geometry. This approach was inspired by successes in the medical field, but because there are essentially an infinite number of features that can be derived using computational geometry, this approach

presents a particular challenge for any machine-learning approach, namely that of input attribute explosion. For example, the approach used to generate the results cited herein uses a field instrument with four (4) time gates and generated 633 attributes based on raw data, statistical properties, and insight from the physics-based MEC discrimination equations (though not the specific inputs from inverse physics model fitting). The geometric attributes are based on finding an optimized ellipsoid that is constructed either automatically using the Lipschitz Global Optimization (LGO) technique [8] or by an expert geophysicist who draws a polygon around the target of interest. To generate the features, the ellipsoid is divided into slices and the features are computed as a whole geometric shape, within quadrants and within the segmentations. Fig. 4 illustrates the computational geometric process of segregating an ellipsoid fit to field data for attribute derivation.

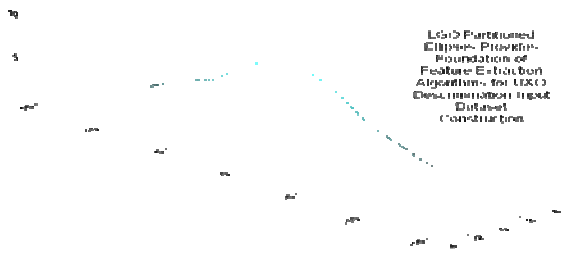

Fig. 4. Development of computational geometrically derived attributes using a globally optimized ellipsoid

Given the prospect that the next generation of geophysical instruments will produce even more data and resultant features, the industry would benefit from an efficient and reproducible site-specific feature reduction methodology – which is precisely the role the information-theoretic approach Minimum Redundancy Maximum Relevance (MRMR) would serve.

3. Attribute Analysis: Computational Geometry and Inverse Physics Models

Our hypothesis is that when the attributes from computational geometry and fitted inverse physics-based modeling approach are combined, the resulting model generated with machine learning will perform better than – or at least as well as – either approach used alone. We will now test this hypothesis first theoretically using information theory, and then empirically, using machine learning.

3.1. Mutual Information Analysis

Understandability of the individual attributes and relationships used for MEC classification analysis is important for the users of the solution. While the computational geometric approach has been shown to be a viable approach, the amount of attributes can make the solution daunting to understand. Methods for feature compression such as principal components analysis, while quite valuable for reducing the number of inputs in a data set used for machine learning, require complex computations to be performed that combine many attributes into a single input vector. This, however, is something that obfuscates solution understandability. In the section below, we describe and test an approach to reduce the attributes required for MEC discrimination modeling using mutual information that offers the additional advantage of preserving the individual attribute identity.

To test the approach on both attribute reduction and relevancy assessment, the data sets from the ESTCP Camp Sibert project [3, 5] are combined so they contain attributes from both the fitted physics-based model parameters and the computational geometric approach; the MEC identity is a binary label (1 for MEC, 0 for non-MEC). The data was collected by others as part of the project and provided to us for this analysis. The EMI data set consists of 174 instances (rows), of which 67 are MEC and 107 are non-MEC. There are seven attributes for the fitted physics-based model and 551 for the computational geometric based model. The MAG data set consists of 182 instances (rows) of which 56 are MEC and 126 are non-MEC. There are six attributes for the fitted physics-based model and 82 for the computational geometric-based model.

Information content has long been used for assessing important of attributes for model building [10]. The method used here is based on mutual information, using a maximum-dependency, minimum-redundancy framework as developed by Peng [7]. This technique provides the necessary theoretical engine to select the best candidate features independent of a machine-learning classifier. The computations are based on the following model:

Given two random variables (x, y), their mutual information (x, y) is defined in terms of their marginal and joint probability density functions $p(x)$, $p(y)$ and $p(x, y)$:

$$I(x,y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dx dy. \quad (2)$$

In terms of mutual information, the goal of feature selection is to develop the set S of m features $\{x_i, \ i = 1...m\}$ which jointly have the largest dependency (or in this case relevance) on the target class, that is the classification of MEC (aka UXO):

$$\max D(S, uxo), where D = \frac{1}{|S|} \sum_{xi \in S} I(xi, uxo). \quad (3)$$

It is likely that using just this formulation will generate a list of features that are redundant with respect to one another (i.e., not all are needed for the same discrimination accuracy); hence, a feature redundancy protective measure is used via a maximum relevance and minimum redundancy formulation:

$$R = \min R(S), \quad R = \frac{1}{|S|^2} \sum_{xi,xj \in S} I(xi, xj). \quad (4)$$

To optimize D (dependency) and R (redundancy) simultaneously, we can use the objective function:

$$\max \Phi(D, R), \ \ \Phi = D - R. \quad (5)$$

The goal of this "maximum relevance, minimum redundancy" (MRMR) approach [7] is to reduce the attribute space. Using a smaller input data set (with the same information content) will result in faster running as well as higher accuracy of machine-learned models. We use it to assess the relative importance/redundancy between the fitted attributes from the physics models and the computational geometry attributes on each of the subsets of the EMI and MAG data.

3.2. Application of MRMR

The first step in applying MRMR to the feature value quantification for the MEC discrimination challenge is preparing the data set. The target of Interest (TOI) is discrete; each case is labeled either as a 1 for MEC or as a 0 for not-MEC. However, the computational geometric paradigm generates features that are represented as continuous variables. Mutual information of discrete variables was used and the variables discritized by using two thresholds: the mean (+/-) alpha*standard deviation as discussed in [7]. MRMR is available as open source, as a web-based application, C and Matlab code.

Table 1. Parameter settings for the MRMR algorithm

MRMR Parameter	Parameter Value Used
Alpha	1.0
Variable states	3
Number variables retained	50
Feature Selection Scheme	Mutual Information Difference (MID)

3.3. MRMR Results

The MRMR ranking produces a rank-ordered list of features, with the top 50 of the 633 candidate EMI features being retained. To understand how to use the MRMR results, consider an example of the top three variables, V1, V2 and V3. This ranking means that if a single variable is desired, then variable V1 should be used. If two variables are desired, then the combination of V1 and V2 is better than the combination of V1 and V3, or V2 and V3. The results discussed below indicate that information contained by developing a feature data set using the computational geometric approach for the EMI data set contains all the information that is contained in the inverse physics modeling. However, the results of MRMR analysis on the MAG data set clearly show the synergy possible when both using both methods are used. This finding is reinforced by the empirical testing via machine learning, as discussed below.

3.4. Investigation of MRMR Results

Since the MRMR analysis resulted in minimal to no selection of the inverse physics model attributes, and these very attributes are what the industry relies on for MEC discrimination, analysis was conducted to further understand this finding.

3.4.1. EMI MRMR Results Analysis

Analysis of the EMI data revealed that the only fitted physics-derived variable in the top 50 rank-ordered set was "Chi^2," which was ranked 45th out of 50 for variable importance. Forty-nine (49) of the features in the top 50 were computational geometric (CG) features. The eigenvalue attributes as described by the inverse physics modeling ($\beta1$, $\beta2$, $\beta3$) – did not appear in the list of top 50 features. Table 2 shows an abbreviated list of features output by the MRMR code.

Table 2. Abbreviated output of MRMR analysis of the EMI data

Feature (attribute) Ranking	Feature (attribute)
#1	CG-500
#2	CG-485
#3	CG-335
…	…
#45	Chi^2

To investigate why the eigenvalue attributes were not ranked with higher priority, we tested whether or not the computational geometric attributes and the eigenvalues were information content redundant. The results of our analysis show that they are, in fact, redundant. To assess to the extent of the redundancy, we developed a function using a common set of eight attributes from the computational geometric data set that explains more than 99% of the variation in each of the eigenvalues $\beta1$, $\beta2$, $\beta3$. Hence, the features developed as part of the computational geometric attribute formulation contain all the information that the eigenvalues have to offer. This is demonstrated via a regression analysis using Multivariate Adaptive Regression Splines (MARS) with 10 times cross-validation [9]. Thus, the need to develop attributes by fitting physics models to the field data is unnecessary, at least in this example. Since the computational geometric approach performs at lower S/N than the inverse physics modeling (10 vs. 100, respectively), more TOI can be discriminated using this method. Moreover, the features that form the inputs to the regression models are those that one would expect such as peak values, ratios between the channels and parameters of power law fits. The results of the computational geometric attribute data set's ability to reproduce all of the fitted physics-based derived attributes are shown in Table 3: R^2 denotes the correlation coefficient.

Table 3. Reproducibility of inversion physics-based EMI features using CG-EMI features

Physics-based parameters obtained by inversion	R^2 obtained using 10 times cross-validation
β1	0,99313
β2	0,99315
β3	0,99320
Chi^2	0,89701
Size	0,98115
Depth	0,99367
Coh	0,76570

In hindsight, it is not surprising that the computational geometric approach includes all of the information that could be available by fitting physics models to the data. After all, we developed the computational geometric model with the discrimination physics in mind. However, this is the first formal analysis that indicates that this information inclusivity is indeed the case. Moreover, these results show that the computational geometric approach can be used to develop a physics-based representation from the EMI Data. Interestingly, the one (Chi^2 a measure of fitness of the inverse fitted-physics model) attribute that did appear in the top 50 features is a solid indicator of how well the inverse physics model is expected to fit the data. It is also important to note that the computation geometric approach was able capture 89,7% of the variation in the expected fitness of the inverse modeling. This ability to predict a priori how an inverse modeling task should perform is extremely valuable for quality assurance/quality control purposes.

3.4.2. MAG MRMR Results Analysis

Analysis of the MAG data revealed three of the physics-derived variables in the top 50 of the rank-ordered set; these are Fit_size (rank #1), Fit_inc (rank #6), and Fit_Depth (rank #48). The remaining 47 of the features in the top 50 were computational geometric attributes. A test of the ability to produce the fitted physics-based attributes from the computational geometric attributes was conducted, this time with very different results as shown in Table 4.

Table 4. Reproducibility of fitted physics-based MAG features using computational geometric MAG attributes

Physics-based parameters obtained by inversion (# is the parameter ranking)	R^2 obtained using 10 times cross-validation
Depth (#48)	0,67
Size (#1)	0,73
Dec	0,21
Inc (#6)	0,49
Solid Angle	0,30
Magnetic Moment	0,49

Clearly, the less well-developed computational geometric approach for MAG sensors is currently not as effective as the EMI approach in capturing the information content from the fitted physics-based inversion model; therefore, further work in this area is warranted.

3.5. MRMR Analysis Summary

This MRMR approach is particularly valuable because it provides gives the ability to screen important features and reject ones of lesser value or that are redundant to making classification predictions without the need to run classification algorithms. This means that important variables can be identified in minutes as opposed to hours or days of simulation computation time. Thus the benefits associated with the machine-learning, algorithm-independent analysis of feature contri-

bution made possible with the MRMR approach are multifold. Not only is it fast and cost-efficient, it guides when easily computed data-driven features should replace more complex ones to obtain features such as those arrived at via fitted physics-based inversion. Additionally, it provides a very fast and efficient screening mechanism to rank the value of new or proposed features, especially when compared to existing features sets. Additionally, these characteristics of the information-theoretic MRMR approach, when corroborated with results from machine-learning algorithms, effectively streamline the understanding of attribute importance and help to focus new research into less well-understood areas. This benefit is discussed in more detail below.

4. Machine Learning Analysis and Results

Machine-learning (ML) techniques are tools that interrogate the information content in the data set and then replace that content with a representative relation(s). That representation can then be used to make predictions relative to unseen instances: in this case sensor data returned from a geophysical investigation.

Based on the information-theoretic MRMR analysis outlined and demonstrated above, we can anticipate and expect certain outcomes when building models from the data sets using machine-learning algorithms and various combinations of attributes. For example, models produced using the EMI data set should rank as:

- Best: Combined Geometric and Fitted Model attributes;
- Second: Geometric attributes, and;
- Third: Fitted physics-attributes.

This ranking reflects the fact that the computational geometric approach replicated the information content in the inverted fitted physics models. The machine-learned model based on the combined geometric-fitted physics attribute data may be slightly better (or tie with) the geometric attribute model, since only one physics attribute appeared in the top 50 features (the measure of the inverse physics model fitness) and then at a very low rank (#45). The data based on the fitted physics models will rank as third accurate, to the extent that it does not contain the information content that the geometric data set provides.

Models produced using the MAG data set are a different story. Clearly, the geometric attributes present valuable information, as do the fitted physics-inversion attributes. One can only conclude, therefore, that the combined CG-physics data set will produce a more accurate model than either data or physics alone.

4.1. Empirical Testing using Machine Learning

The models were constructed from the EMI and MAG data sets (fitted physics, geometric, geometric-fitted physics). All models were developed using 10 times cross-validation, and all used the designated technique subset (not just the subset of the top 50 features identified above). The tool used was TreeNET [9] and used with default settings, except the number of trees was set to 2,000.

4.1.1. EMI Machine Learning Results

The model results using the EMI inverse physics model data set is provided in fig. 5 and show respectable MEC discrimination (ROC>0,95). A receiver operating characteristic (ROC) Chart is one that plots the accuracy of a classifier over the data set; true positive rate of detection on the y-axis and the false positive rate on the x-axis. The area under ROC curve (AUC) is used as a measure of quality of a probabilistic classifier, with an area of 1.0 being best achievable, and 0.50 (blue line) being no better than random guessing. The graph shows the order of MEC removal, progressing from left to right, with the final excavation occurring at the right-most section of the

graph. The last remaining MEC item is removed when the value of the y-axis is 1.0. For MEC removal projects, additional MEC would be removed beyond the last known MEC item as a means of validation and stakeholder acceptance.

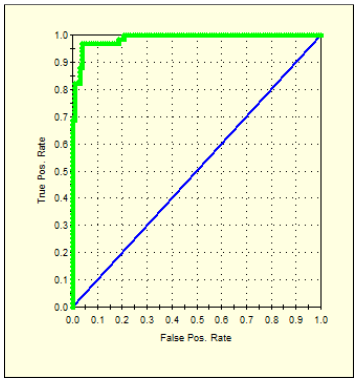

Fig. 5. EMI: Inverse Physics Model, ROC (AUC) =0,98786

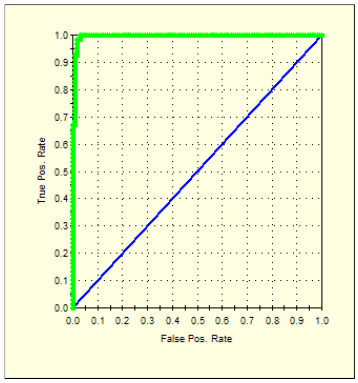

Fig. 6. EMI: Computational Geometry, ROC (AUC) =0,99609

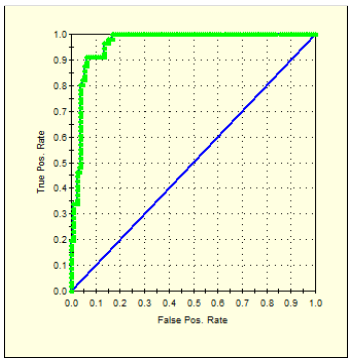

Fig. 7. MAG: Inverse Physics Model, ROC (AUC) =0,96358

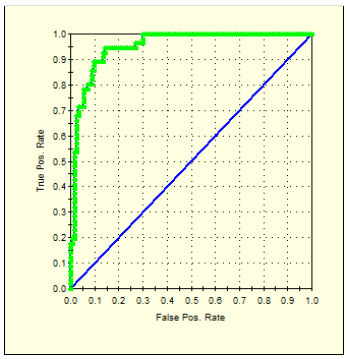

Fig. 8. MAG: Computational Geometry, ROC (AUC) =0,95366

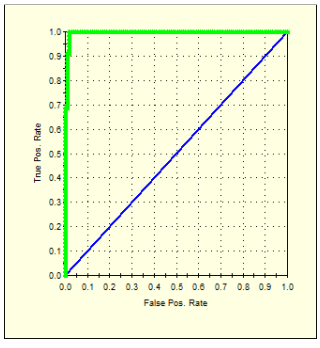

Fig. 9. EMI: Computational Geometry and Inverse Physics Model, ROC (Are AUC)=0,99623

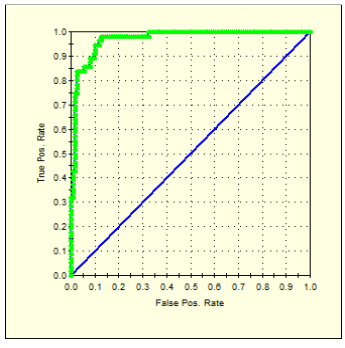

Fig. 10. MAG: Computational Geometry and Inverse Physics Model, ROC (AUC) =0,97251

The model results using the EMI geometric data set (provided in fig. 6) also show respectable MEC discrimination (ROC > 0,95) with a much better time/speed curve for identifying MEC.

4.1.2. MAG Machine-Learning Results

The model results using the MAG fitted physics data set, provided in fig. 7, show respectable MEC discrimination (ROC > 0,95).

The model results using the MAG geometric data set provided in fig. 8 also show respectable MEC discrimination (ROC > 0,95), but with a slightly worse curve indicating slower identification of the final MEC item.

4.1.3. Combined EMI and MAG Analysis

EMI: The model results using the EMI-geometric and inverse physics model data set provided in fig. 9 also show respectable MEC discrimination (ROC>0,95), again with a slightly better curve showing faster classification of MEC signals. The expectation of the classifier performance is in concert with the understanding gained from the information-theoretic MRMR analysis. The computational geometric model performed better than the inverse physics model, because it replicated basically all the important the information content of the fitted physics model and also generated additional information essential for higher accuracy MEC classification. The combined geometric-fitted physics model slightly outperformed the geometric-only model; this model included the fitness of the inverse model to the data.

MAG: The model results using the MAG geometric and inverse physics model data set (shown in fig. 10) also demonstrate respectable MEC discrimination (ROC>0,95), with a higher AUC ROC value with a slightly better curve, but again indicate slower identification of the final MEC found. The expectations of the classifier performance are in concert with the information-theoretic MRMR analysis in terms of overall performance (a higher AUC ROC value was obtained using the combined data-physics data sets). However, the overall identification of that last MEC was slower; hence this solution would require more holes to be dug (and non-MEC items excavated) than the other ones. These mixed results are indicative of a less than fully developed MAG data and inverse-physics model.

5. Over-Fitting Test

In machine learning, over-fitting (also known as "memorizing") is an important issue with respect to assuring predictable performance on unseen data. Being able to predict how an algorithm will do on unseen data is more important than the algorithm doing well on training data. To guard against over-fitting, large data sets are divided into training, testing, and/or validation subsets (where the model's performance is solely judged on the validation performance statistics) or as in our case, into a larger number of smaller data sets where a 10 times cross-validation approach could be used.

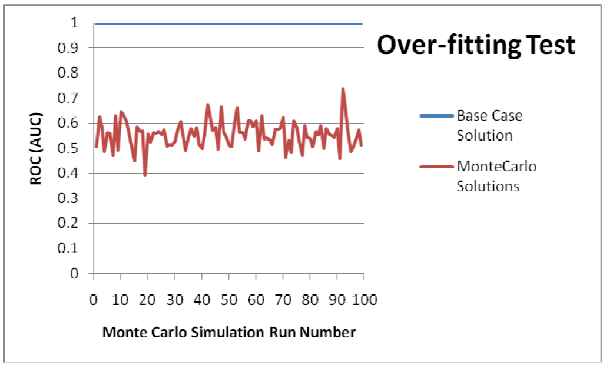

Fig. 11. Machine learning over-fitting test using a Monte-Carlo target (aka label) column scrambling

The method used in this study to test for over-fitting is to scramble the target columns of the EMI-geometric attribute data set using a Monte Carlo method. The target variable (MEC or non-MEC) is solved once; then the target column is scrambled 99 times (for a total of 100 runs) using the exact same parameter settings and a machine-learning model is then again built. If there is a structural flaw in the experiment, it will show up. Specifically, if the accuracy of the solutions developed with the Monte Carlo-scrambled (randomized labels) targets are similar to the true target sequencing; this is not a good result. This technique provides qa/qc check and safe-guards against deploying models that, while they may look good in testing phase, they are but mere chance findings and fitting noise – likely to perform poorly upon deployment. As shown in fig. 11, the true solution (AUC=0,997) exceeds both the average (AUC=0,556, STDEV=0,054) and maximum (AUC=0,737) of the Monte Carlo over-fitting test. Our conclusion is that the modeling approach is valid, the model is identifying a signal (not just fitting noise); therefore, the results are expected to be reasonably reliable for their intended purpose.

6. Summary and Results

We demonstrate the value and understandability of the computational geometric MEC discrimination method, and developed a methodology for understanding the value of MEC features by applying information theory. We used machine learning to fuse the information content of attributes derived from both machine-learning computational geometric and from fitted physics-based models.

The authors believe that the inverse physics modeling, while providing great insight, over compresses the information available in the geophysical signals into too few variables and hence impose an artificial limit on that methods accuracy. The multi-disciplinary computational geometric (MDCG) is intended to extend – not replace-this deep physics-based understanding by sup-

plementing the discrimination information with factors the inverse physics modeling approach cannot capture. For the test case, the MDCG approach was found to contain all the information (in eight common variables) contained in the EMI data set, but not the MAG data set.

The empirical tests conducted using machine-learning are consistent with their performance predicted using information theory. We further identified the information overlap between our computational geometric approach and the fitted physics model approach by others: complete overlap for the EMI sensor – indicating a rational physical basis for the method – and a partial overlap for the MAG sensors.

REFERENCES

1. Bell T.H. Subsurface Discrimination Using Electromagnetic Sensors / T.H. Bell, B.J. Barrow, J.T. Miller // IEEE Transactions on Geoscience and Remote Sensing. – 2001. – Vol. 39, N 6. – P. 1286 – 1293.
2. Using Machine Learning to Compliment and Extend the Accuracy of UXO Discrimination Beyond the Best Reported Results of the Jefferson Proving Ground / L.M. Deschaine, R.A. Hoover, J. N. Skibinski [et al.] // Technology Demonstration. Society for Modeling and Simulation International's Advanced Technology Simulation Conference. – San Diego, 2002. – April. – P. 46 – 52.
3. Deschaine L. M. Advanced MEC Discrimination Comparative Study on Standardized Test-Site Data Using Linear Genetic Programming (LGP) Discrimination (MM-0811) [Електронний ресурс] / L.M. Deschaine. – Completed 2009. – Режим доступу: http://www.estcp.org/Technology/MM-0811-FS.cfm.
4. ESTCP. Technical Report Description and Features of UX-Analyze ESTCP Project MM-0210. – 2009. – 42 p.
5. Keiswetter D. SAIC Analysis of Survey Data Acquired at Camp Sibert / D. Keiswetter // Interim Report, ESTCP Project MM-0210. – 2008. – July. – P. 112.
6. Nordin J.P. A Compiling Genetic Programming System that Directly Manipulates the Machine Code / J.P. Nordin // Advances in Genetic Programming / K. Kinnear (ed.). – MIT Press, Cambridge MA, 1994. – P. 311 – 331.
7. Peng H. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy / H. Peng, F. Long, C. Ding // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2005. – Vol. 27, N 8. – P. 1226 – 1238.
8. Pintér J. D. Global Optimization in Action / J. D. Pintér // Kluwer Academic Publishers, Dordrecht. Now distributed by Springer Science and Business Media. – New York, 1996. – 512 p.
9. Salford Systems Inc. Salford Data Miner Users Manual: CART Version 6.4, TreeNET Version 2.0, MARS Version 3.0, and Random Forrest Version 1.0. – San Diego, CA, 2009.
10. Varmuza K. Monatsh. Chem / K. Varmuza. – 1974. – Vol. 105, N 1.

Стаття надійшла до редакції 28.05.2010