MULTI-LABEL CLASSIFIER PERFORMANCE EVALUATION WITH CONFUSION MATRIX

Damir Krstinić, Maja Braović, Ljiljana Šerić and Dunja Božić-Štulić

Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, University of Split, R. Boškovića 32, Split 21000, Croatia

ABSTRACT

Confusion matrix is a useful and comprehensive presentation of the classifier performance. It is commonly used in the evaluation of multi-class, single-label classification models, where each data instance can belong to just one class at any given point in time. However, the real world is rarely unambiguous and hard classification of data instance to a single class, i.e. defining its properties with single distinctive feature, is not always possible. For example, an image can contain multiple objects and regions which makes multi-class classification inappropriate to describe its content. Proposed solutions to this set of problems are based on multi-label classification model where each data instance is assigned one or more labels describing its features. While most of the evaluation measures used to evaluate single-label classifier can be adapted to a multi-label classification model, presentation and evaluation of the obtained results using standard confusion matrices cannot be expanded to this case.

In this paper we propose a novel method for the computation of a confusion matrix for multi-label classification. The proposed algorithm overcomes the limitations of the existing approaches in modeling relations between the classifier output and the Ground Truth (i.e. hand-labeled) classification, and due to its versatility can be used in many different research fields.

KEYWORDS

Classification, multi label classifier, performance evaluation, confusion matrix

1. INTRODUCTION

Multi-class classification (MCC), where each data instance or object is assigned to a class from the set of a priori known classes, is widely encountered in scientific literature and engineering applications. Regardless of the number of possible classes, they are mutually exclusive and each object can be assigned to only one class. This approach in machine learning, also known as Single-label classification [18], relies on fundamental assumption that each data object belongs to only one concept and has a unique semantic meaning [21]. While this approach is well-known and widely used in supervised learning, there are data sets that are too complex to impose the restriction of only one label for each data instance [5], [12]. These problems include text categorization, sentiment and emotion recognition, semantic scene classification and many other problems. In fact, most of real life data is often difficult to describe with a single distinctive feature. Objects, phenomena, relations, interactions and other manifestations of natural and artificial processes are seldom simple enough to be easily defined and unambiguously classified. Machine learning approaches dealing with data that

David C. Wyld et al. (Eds): SAIM, SIPM, ACSIT, FCST, ICITE - 2020
pp. 01-14, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.100801

cannot be simply classified to a single distinctive class, i.e. described with a unique semantic context, are referred to as Multi-label classification (MLC) [18], [17] or Multi-label learning [5], [19] gain the attention and becomes a relevant research area [1], [21].

Beside assigning more than one semantic concept to each data instance, MLC algorithms have to deal with other difficulties, e.g. correlation between different labels and an uneven number of label occurrences on the data. Representative example is image classification. Aside from rare exceptions, most of the real life images contain multiple objects and regions with their interactions and attributes and can be annotated with multiple labels [19]. Some labels can be very common in the data set and appear in almost all images (e.g. Sky), while others could be rare and appear only on a few images in the whole data set. Moreover, it is not uncommon for different labels to look very similar in the image context (e.g. Clouds, Smoke or Fog), making hard even for a human to create unambiguous Ground Truth (GT) labeling [2]. For a researcher working on a MLC algorithm, it is essential to have a tool that does not only evaluate algorithm effectiveness but can also reveal relations between labels and clearly indicate the weaknesses of the classifier.

Confusion matrices have been present in the evaluation of scientific models and engineering applications for a long time and are commonly used in many different areas such as computer vision [14], natural language processing [10], acoustics [16], etc. In its simplest form a confusion matrix shows a binary classifier performance in table with two rows and two columns [3], [13], [4] and represents the percentages of four possible classification outcomes: True Positive (TP), False Positive (FP), True Negative (TN) and False negative (FN). This principle is easily extended to visualization of results obtained by Multi-class classification model [11], where each object form the data set can belong to just one of many distinctive classes at any given point in time. If, for example, an object of type A is often misclassified as type B, the confusion matrix will clearly reveal this and suggest to a researcher that she or he could improve the classification model by looking for additional features that can help better distinguish classes A and B.

Even though confusion matrices are adequate for the visualization of results obtained by MCC models, they fail when it comes to Multi-label classification where an object from a data set can simultaneously belong to multiple classes. Since the analysis of the confusion matrix could provide insight into the relations between different data features and objects and also reveal inherent structure of the data itself, it is important to find a way in which a confusion matrix can be applied for evaluation of MLC models. In this paper we propose a method for representation of Multi-label classification results with confusion matrices.

The rest of the paper is structured as follows. Multi-label classification paradigm and existing evaluation measures for Multi-label classification model are presented in section II. In section III a novel approach for the computation of confusion matrices for Multi-label classification problems is proposed and elaborated in details. In section IV we give a conclusion and discuss future work.

2. MULTI-LABEL CLASSIFIER PERFORMANCE EVALUATION

Multi-label classification is a supervised learning paradigm where each data instance can be assigned more than one label from the set of predefined labels. This approach gained a lot of attention in recent years as it is applicable whenever the data set is too complex to assign each data instance to a single distinctive class, i.e. characterize each data instance with a single distinctive feature or concept. Reported applications of multi-label classification include [17], [5] text categorization,

image classification, graph classification, bioinformatics, gene function analysis, emotion and sentiment recognition, multimedia annotation, social network analysis and many more. Surveys of multi-label classification techniques and state of the art approaches are given in [18], [17], [5], [21].

We define Multi-label classification problem according to [21], [5]:

Let $\mathcal{X} = R^d$ be an input space of d-dimensional data instances and $\mathcal{Y} = \{\lambda_1, \ldots, \lambda_q\}$ the output label space with $|\mathcal{Y}| = q$ possible labels that can be assigned to each data instance. Multi-label pattern is a pair (x, Y) where $x \in \mathcal{X}$ is a data instance and $Y \subseteq \mathcal{Y}$ is a set of associated true labels. Label set Y is represented as a q dimensional binary vector $Y \in \{0,1\}^q$ where labels relevant to x are represented by 1 and labels irrelevant to x are represented by 0.

The task of multi-label training is to learn function $\mathcal{H}_{\mathcal{ML}}: \mathcal{X} \to 2^{|\mathcal{Y}|}$ which predicts a set $Z \subseteq \mathcal{Y}$ of relevant labels for an unseen data instance. Note that Multi-class classification could be defined as a special case of Multi-label classification where $\mathcal{H}_{\mathcal{MC}}: \mathcal{X} \to \mathcal{Y}$ predicts a single class associated to data instance [5].

2.1. Evaluation Measures

Evaluation measures used to evaluate performance of Multi-class classifier are usually based on hit and miss ratio on an unseen test data with associated Ground Truth classes. Prediction of the classifier $\mathcal{H}_{\mathcal{MC}}: \mathcal{X} \to \mathcal{Y}$is accurate only if the predicted class is the same as the GT class. In Multi-label classification, prediction could be completely accurate if predicted labels Z are exactly the same as GT labels Y , partially accurate if $Y \cap Z \neq \emptyset$ or completely inacurate if $Y \cap Z = \emptyset$. Thorough survey of the evaluation techniques for Multilabel classification with correlation analysis of different performance measures is given in [1].

Let $D_t = \{(x_i, Y_i) | i = 1, \ldots, N\}$ be a set of multi-label patterns where Y_i is the Ground Truth set of labels for data instance x_i unseen by the classifier $\mathcal{H}_{\mathcal{ML}}: \mathcal{X} \to 2^{|\mathcal{Y}|}$and $Z_i = \mathcal{H}_{\mathcal{ML}}(x_i)$ is the prediction of the classifier. Evaluation measures for evaluating performance of the MLC classifier are divided into example-based metrics and label-based metrics. Example-based metrics calculate performance for every data instance and average over the entire data set, while label-based metrics evaluate performance for each label individually and then average across all labels.

2.1.1. Example-based evaluation measures

Example-based evaluation measures are calculated by taking into account each instance hit and miss ratio regardless of label and averaging over the entire test set. Example based Accuracy, Precision and Recall are define with [1], [6]:

$$Accuracy_{EB}(\mathcal{H}_{\mathcal{ML}}, \mathcal{D}_t) = \frac{1}{N} \sum_{i=1}^{N} \frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \quad (1)$$

$$Precision_{EB}(\mathcal{H}_{\mathcal{ML}}, \mathcal{D}_t) = \frac{1}{N} \sum_{i=1}^{N} \frac{|Y_i \cap Z_i|}{|Z_i|} \quad (2)$$

$$Recall_{EB}(\mathcal{H}_{\mathcal{ML}}, \mathcal{D}_t) = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap Z_i|}{|Y_i|} \quad (3)$$

Using example-based Precision and Recall, F_1score [7] can be computed, representing the weighted average between Precision and Recall [1]:

$$F_{1EB} = \frac{1}{N}\sum_{i=1}^{N}\frac{2|Y_i \cap Z_i|}{|Y_i| + |Z_i|} \quad (4)$$

Hamming loss:

$$HammingLoss(\mathcal{H}_{\mathcal{ML}}, \mathcal{D}_t) = \frac{1}{N}\sum_{\mathrm{i=1}}^{\mathrm{N}}\frac{|Y_i \,\Delta\, Z_i|}{|L|} \quad (5)$$

where Δ stands for a symmetric difference of two sets, which is equivalent to the XOR operation in Boolean logic [1]. Hamming Loss (5) is a widely used evaluation measure for MLC, penalizing difference between predicted and GT labels. Both labels that are predicted and do not exist in GT and labels that are not predicted but exist in GT are taken into the account.

Subset Accuracy:

$$SubsetAccuracy(\mathcal{H}_{\mathcal{ML}}, \mathcal{D}_t) = \frac{1}{N}\sum_{\mathrm{i=1}}^{\mathrm{N}}[\![Y_i = Z_i]\!] \quad (6)$$

where $[\![\Delta]\!]$ is Iverson bracket [20], mapping true logic condition to 1 and false to 0. Subset Accuracy or Exact Match is a rigid measure considering prediction accurate only if it is exactly the same as GT.

2.1.2. Label-based evaluation measures

Label-based evaluation considers every label separately, reducing Multi-label classifier to a binary classifier for a particular label, with four possible prediction outcomes: True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). Accuracy, Precision and Recall are defined by:

$$Accuracy = \frac{TP + TN}{N}$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 = 2\frac{Precision \cdot Recall}{Precision + Recall}$$

Label-based $F_1 score$ is defined by:

$$F_1 = 2\frac{Precision \cdot Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \quad (8)$$

Label-based classification metrics for the classifier $\mathcal{H}_{\mathcal{ML}}$and dataset $\mathcal{D}_t$ could be obtained by using macro or micro averaging techinques. Let B be any of the measures defined by equation (7). $B_{macro}(\mathcal{H}_{\mathcal{ML}}, \mathcal{D}_t)$ and $B_{micro}(\mathcal{H}_{\mathcal{ML}}, \mathcal{D}_t)$are calculated by [1]:

$$B_{macro}(\mathcal{H}_{\mathcal{ML}}, \mathcal{D}_t) = \frac{1}{q}\sum_{j=1}^{q} B\left(TP_j, FP_j, TN_j, FN_j\right) \quad (9)$$

$$B_{micro}(\mathcal{H}_{\mathcal{ML}}, \mathcal{D}_t) = B\left(\sum_{j=1}^{q} TP_j, \sum_{j=1}^{q} FP_j, \sum_{j=1}^{q} TN_j, \sum_{j=1}^{q} FN_j\right) \quad (10)$$

3. MULTI-LABEL CLASSIFICATION CONFUSION MATRIX

Multi-label classification measures presented in the previous section are widely used in scientific papers. Detailed literature reviews on evaluation techniques for MLC are given in [1], [15]. While wide variety of proposed measures can be used to evaluate performance of a MLC algorithm, there is little information on what is happening with data instances which are labeled inaccurately. If, for example, label $\lambda \in \mathcal{Y}$ has poor Recall, i.e. if of λ is usually not assigned to data instances for which it is relevant, information on what labels are often assigned instead of λ could be very useful. In order to optimize classifier performance it is crucial to gain deeper insight into the internal classifier operations and the relations amongst the different labels. This information could also be used to select new discriminative features on the data set.

3.1. Multi-Class Confusion Matrix

In the case of Multi-class classification (MCC), where each data instance is assigned to a single distinctive class, Confusion matrix is a useful and comprehensive presentation of the classifier performance on a data set with known true labels. Moreover, most of the evaluation metrics can be represented as a function of the Confusion matrix entries [12].

For the Multi-class classifier where $\mathcal{H}_{\mathcal{MC}}: \mathcal{X} \to \mathcal{Y}$predicts a single class, Confusion matrix is constructed by comparing the predicted class with the known GT class [8]. Each row of the matrix represents true label (GT) and each column of the matrix represents the prediction of the classifier$\mathcal{H}_{\mathcal{MC}}$. For each data instance x with GT class Y and predicted class Z, matrix cell corresponding to the Y -th row and Z-th column is incremented, counting the number of times that the object of class Y is assigned to a class Z. This way raw Confusion matrix is constructed where diagonal elements of the matrix represent the number of accurate classifications for each class, while off-diagonal elements represent missclassifications. Precision and Recall for each class separately

can be directly computed from the raw Confusion matrix. Precision for each class is computed by dividing a diagonal element of the Confusion matrix with the sum of all elements in the corresponding column. Recall for a class is computed by dividing a diagonal element of the matrix by the sum of all elements in the row.

An example of the Confusion matrix for MCC data set with $N = 45$ data instances and $q = 4$ classes is shown in Table I. Precision for each class is given in the last column, while Recall value for each class is given in the last row.

Table 1. Example of the confusion matrix for four class multi-class classifier. Precision for each class is shown in the last row. Recall for each class is shown in last column.

		Predicted class				
		λ_1	λ_2	λ_3	λ_4	Recall
GT class	λ_1	8	0	0	0	1.00
	λ_2	4	9	1	1	0.60
	λ_3	3	0	7	0	0.80
	λ_4	1	0	2	9	0.75
Precision		0.50	1.00	0.70	0.90	

Table 2. Precision matrix computed from the confusion matrix given in Table 1.

	λ_1	λ_2	λ_3	λ_4
λ_1	0.5	0	0	0
λ_2	0.25	1	0.1	0.1
λ_3	0.19	0	0.7	0
λ_4	0.06	0	0.2	0.9

Table 3. Recall matrix computed from the confusion matrix given in Table 1.

	λ_1	λ_2	λ_3	λ_4
λ_1	1	0	0	0
λ_2	0.27	0.6	0.07	0.07
λ_3	0.3	0	0.7	0
λ_4	0.08	0	0.17	0.75

Confusion matrix normalization provides further information on relations between classes and types of classification errors [9]. Recall matrix is computed by dividing each cell of the raw Confusion matrix by the sum of all elements in the corresponding row. Diagonal elements of the Recall matrix are Recall values computed for each class, given in the last row of the raw Confusion matrix (Table I). Off-diagonal elements for the row representing true class Y represent the probability that the object of the class Y will be misclassified as class Z, where $Y \neq Z$. Precision matrix is computed by dividing each cell of the raw Confusion matrix with the sum of the corresponding column. Diagonal elements of the Precision matrix represent precision for the corresponding class given in the last

column of the raw Confusion matrix. Off-diagonal elements of a column representing class Z are probabilities that the object assigned to class Z really belongs to class Y , $Y \neq Z$. Precision matrix computed from the raw Confusion matrix is given in Table 2 and Recall matrix is shown in Table 3.

3.2. Confusion-Matrix in MLC Paradigm

As illustrated in the previous example, a Confusion matrix provides comprehensive insight into the classifier performance for the MCC problems. However, there are several obstacles for the extension of this simple yet effective principle to the Multi-label classification paradigm. The contribution of the data instance x to the raw Confusion matrix is straightforward only in the trivial case where $|Y| = |Z| = 1$. If either Y or Z has more than one label the situation is not so clear. In the following paragraphs we will propose the algorithm for the computation of Confusion-matrices for Multi-label classification. In this work we assume that MLC predicts at least one label for each data instance and cardinality of both Y and Z is greater than 0.

Let us consider four possible scenarios for sets of true labels Y and predicted labels Z:

(i) GT and predicted labels for data instance x are exactly the same, $Y = Z$, i.e. the classifier accurately predicts relevant labels. Contribution C for data instance x to the Confusion matrix is accounted by incrementing diagonal elements corresponding to label set Y:

$$C = diag(Y) \qquad (11)$$

(ii) Classifier prediction and GT differ. Prediction Z contains labels that are not relevant to data instance x. True label set Y does not contain labels that do not exist in prediction Z, i.e. all relevant labels are predicted by the classifier:

$$|Y \setminus Z| = 0, \quad |Z \setminus Y| > 0, \qquad (12)$$

where $Y \setminus Z$ represents a set of labels that exist in GT and do not exist in predicted set Z, and $Z \setminus Y$ represents a set of labels that are predicted by the classifier and are not relevant to x. Although all relevant labels from Y are accurately predicted, the contribution of data instance x can not be accounted for by simply incrementing the appropriate diagonal elements. It is reasonable to assume that some features of $x \in \mathcal{X}$ corresponding to concepts and semantic meaning connected with true labels Y misslead the classifier to predict non-existing labels. Proportional share from each true label contribution should be accounted for to the labels that exist in prediction and do not exist in GT. Contribution C of the data instance x to the Confusion matrix is:

$$C = [Y \otimes (Z \setminus Y) + |Y| \cdot diag(Y)\,]/|Z| \qquad (13)$$

First element in square brackets is the outer product that redistributes the contribution of true labels to inaccurately predicted labels, i.e. accounts part of contribution of each label from Y equaly to all labels in Z which are not relevant to x. Second summand in square brackets increments diagonal elements corresponding to labels in Y . Overall contribution is normalized by $|Z|$ to accurately model the distribution of $|Y|$ true labels to contribution C. The sum of each row representing one true label is equal to one, while the sum of all

elements of C is equal to number of true labels $|Y|$. Equation (13) takes equal parts of the contribution from each true label and transfers it to the inaccurately predicted labels.

(iii) Prediction and GT differ. GT contains labels that do not exist in prediction. Prediction does not contain labels that do not exist in GT:

$$|Y\backslash Z| > 0, \quad |Z\backslash Y| = 0 \qquad (14)$$

It is reasonable to assume that some features of x corresponding to true labels Y are not recognized accurately by the classifier and are attributed to other true labels from Y that are predicted in Z. Contribution of the labels that are not accurately predicted should be equally distributed to all labels that exist in both GT and prediction. Contribution to the Confusion matrix is:

$$C = [(Y\backslash Z) \otimes Z]/|Z| + diag(Z) \qquad (15)$$

First summand in the equation (15) redistributes the contribution of relevant labels that are not predicted equally to all predicted labels. Second summand accounts the contribution of the predicted labels to the diagonal elements of the Confusion matrix. The sum of rows corresponding to true labels in Y is one, while sum of all elements in C is equal to $|Y|$.

(iv) Prediction and GT differ. GT contains labels that do not exist in prediction. Prediction also contains labels that do not exist in GT:

$$|Y\backslash Z| > 0, \quad |Z\backslash Y| > 0 \qquad (16)$$

Diagonal elements corresponding to labels that exist in both GT and prediction (if any) should be incremented. We consider these labels accurately classified. Contribution of other labels that exist in GT and do not exist in prediction should be equally distributed among labels that are predicted but do not exist in GT. Contribution to the Confusion matrix is:

$$C = [(Y\backslash Z) \otimes (Z\backslash Y)]/|Z\backslash Y| + diag(Y \cap Z) \qquad (17)$$

First summand equally redistributes contribution of GT labels that are not accurately predicted to predicted labels not relevant to data instance x, normalized by cardinality of $Z\backslash Y$ set. Second summand accounts for the contribution of accurately predicted true labels, if any. The sum of each row corresponding to labels in Y is equal to one. The sum of all elements of C is equal to cardinality of Y .

Based on the previous analysis, the algorithm that computes Multi-label Confusion matrix $\mathcal{M}$ for dataset $D = \{(x_i, Y_i) | \, i = 1, \ldots N\}$, where x_i is data instance and $Y_i \in \{0,1\}^q$ its true labels represented as q-dimensional binary vector, and classifier $\mathcal{H}_{\mathcal{ML}}$ capable of predicting labels $Z_i = \mathcal{H}_{\mathcal{ML}}(x_i), \quad Z_i \in \{0,1\}^q$ for unseen data instances x_i is given in Algorithm 1.

Algorithm 1 Multi-label confusion matrix

Input: $D = \{x\_i, Y\_i) \mid i = 1, \ldots, N\}$
Output: $\mathcal{M}$
$\mathcal{M} \leftarrow zeros(q \times q)$
for $i \leftarrow 1{:}N$ ***do***
$Z_i \leftarrow \mathcal{H}_{\mathcal{ML}}(x_i)$
if $Y_i = Z_i$ ***then***
$C \leftarrow diag(Y_i)$
else
if $|Y_i\ Z_i| = 0$ ***then***
$C \leftarrow [(Y_i \cap Z_i) \otimes (Z_i \setminus Y_i) + |Y_i| \cdot diag(Y_i)]/|Z_i|$
else if $|Z_i \setminus Y_i| = 0$ ***then***
$C \leftarrow [(Y_i \setminus Z_i) \otimes Z_i]/|Z_i| + diag(Z_i)$
else
$C \leftarrow [(Y_i \setminus Z_i) \otimes (Z_i \setminus Y_i)]/|Z_i \setminus Y_i| + diag(Y_i \cap Z_i)$
end if
end if
$\mathcal{M} \leftarrow \mathcal{M} + C$
end for

Algorithm 1 requires single pass through the data set to compute raw Confusion matrix for a Multi-label classifier. Once the confusion matrix is constructed, Precision and Recall matrices can easily be computed by normalizing raw confusion matrix with sum of column elements or sum of row elements, respectively [9].

3.3. Example

Let us illustrate Algorithm 1 by using a simple example. Data set $\mathcal{D}$ with $|\mathcal{D}| = 7$ samples and $q = 4$ possible labels is shown in Table 4. Scenario for computing contribution C for MLC confusion matrix $\mathcal{M}$, according to the analysis given in subsection 3.2 is shown below each data instance x_i. Computed contribution C for each x_i is given in the last row (only diagonal and non-zero elements are shown).

Table 4. Example of multi-label Confusion matrix calculation. Top to bottom: GT and prediction vectors for 7 data instances with 4 possible labels; scenario for computing contribution (see subsection 3.2) contribution of each data instance to Confusion matrix.

	x_1				x_2				x_3				x_4				x_5				x_6				x_7			
X_i	1	1	0	0	0	1	1	0	0	0	0	1	1	1	1	1	0	1	1	0	0	1	1	0	0	1	0	1
Y_i	1	1	0	0	1	1	1	0	1	0	0	1	0	1	1	1	0	1	0	0	1	1	0	0	1	0	1	0
	(i)				*(ii)*				*(ii)*				*(iii)*				*(iii)*				*(iv)*				*(iv)*			
C	1				0				0				0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0				0				0			
		1			$\frac{1}{3}$	$\frac{2}{3}$				0				1				1				1			$\frac{1}{2}$	0	$\frac{1}{2}$	
			0		$\frac{1}{3}$		$\frac{2}{3}$				0				1			1	0		1		0				0	
				0				0	$\frac{1}{2}$			$\frac{1}{2}$				1				0				0	$\frac{1}{2}$		$\frac{1}{2}$	0

Contribution of $x_1$with $Y_1 = Z_1$is straightforward and does not require deeper elaboration. For $x_2$prediction contains two labels that exist in GT, and one additional label that does not exist in GT (scenario (ii): $|Y_2 \backslash Z_2| = 0, |Z_2 \backslash Y_2| > 0$). Contribution is computed according to eq. (13):

$$\begin{aligned} Y_2 \cap Z_2 &= [0 \quad 1 \quad 1 \quad 0] \\ Y_2 \backslash Z_2 &= [0 \quad 0 \quad 0 \quad 0] \\ Z_2 \backslash Y_2 &= [1 \quad 0 \quad 0 \quad 0] \end{aligned} \tag{18}$$

$$\begin{aligned} C &= [(Y_2 \cap Z_2) \otimes (Z_2\ Y_2) + |Y_2| \cdot diag(Y_2)]/|Z_2| \\ &= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & & \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned} \tag{19}$$

Resulting contribution computed in eq. (19) takes into the account that true labels are partially misclassified as non-existent labels and redistributes proportional part of the contribution while keeping sum of all elements of contribution matrix equal to $|Y_2|$. Contribution of the x_3 is computed according to the same scenario.

For x_4GT contains label that is not accurately predicted, i.e.$|Y_4 \setminus Z_4| > 0, |Z_4 \setminus Y_4| = 0$. Contribution is computed according to eq. (15):

$$\begin{aligned} Y_4 \cap Z_4 &= [0 \quad 1 \quad 1 \quad 1] \\ Y_z\ Z_4 &= [1 \quad 0 \quad 0 \quad 0] \\ Z_4 \backslash Y_4 &= [0 \quad 0 \quad 0 \quad 0] \\ C &= [(Y_4 \backslash Z_4) \otimes Z_4]/|Z_4| + diag(Z_4) \end{aligned} \tag{20}$$

$$= \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{21}$$

Equation (21) increments diagonal matrix elements for accurately predicted labels and distributes evenly the contribution of true label that is inaccurately predicted to predicted labels. Contribution of x_5 is computed in the same manner.

Samples x_6 and x_7 represent scenario (iv) where both GT and prediction contain labels that do not exist in the opposite vector, with contribution computed according to the equation (17). For x_6 label that exists in both GT and prediction is considered correctly assigned and corresponding diagonal element is incremented. The contribution of other true label is assigned to inacurrately predicted

label. For$x_7$both true labels are inaccurately classified as predicted labels that do not exist in GT. As it is not possible to conclude which features were misinterpreted, contribution of the labels is evenly distributed across the predicted labels.

Table 5. Raw confusion matrix $\mathcal{M}$ for the example given in Table 4. Precision for each class is shown in the last row, Recall for each class is shown in last column.

		Predicted class				
		λ_1	λ_2	λ_3	λ_4	Recall
GT class	λ_1	**1.00**	0.33	0.33	0.33	0.50
	λ_2	0.83	**4.67**	0.50	0.00	0.78
	λ_3	1.33	1.00	**1.67**	0.00	0.42
	λ_4	1.00	0.00	0.50	**1.50**	0.50
Precision		0.24	0.78	0.56	0.82	

Resulting raw MLC confusion matrix is show in Table 5. Unlike raw multi-class confusion matrix, entries in the raw multi-label confusion matrix are floating point numbers, as contribution of each label in GT could be split amongst labels in prediction. Precision matrix obtained by normalization of raw confusion matrix is shown in Table 6 and Recall matrix is shown in Table 7.

Table 6. Precision matrix computed from the raw confusion matrix $\mathcal{M}$.

	λ_1	λ_2	λ_3	λ_4
λ_1	0.24	0.06	0.11	0.18
λ_2	0.20	0.78	0.17	0
λ_3	0.32	0.17	0.56	0
λ_4	0.24	0	0.17	0.82

Table 7. Recall matrix computed from the raw confusion matrix $\mathcal{M}$.

	λ_1	λ_2	λ_3	λ_4
λ_1	0.50	0. 17	0. 17	0. 17
λ_2	0.14	0.78	0.08	0
λ_3	0.33	0.25	0.42	0
λ_4	0.33	0	0.17	0.50

4. CONCLUSIONS

Confusion matrix is a useful and comprehensive presentation of classifier performance. It is not just another way of computing Precision, Recall or some other evaluation measure, it is rather a magnifier that provides us with deeper insight into the classifier internal operation. Inspection of the confusion matrix and its derivatives provide us with strong clues on relations between classes and labels representing semantic meanings and concepts assigned to data instances. Confusion matrix reveals classifier weaknesses and suggests guidelines for further research and improvement in model

performance. Analysis of the confusion matrix could also provide insight into relations between different data features and objects and reveal inherent structure of the data itself.

In this paper a method for the computation of a confusion matrix for the Multi-label classification model is proposed. The proposed algorithm overcomes the limitations of the existing approaches in modeling relations between the classifier output and the Ground Truth classification. Multi-class classification can be considered as a special case of multi-label classification with imposed limitation $|Y| = |Z| = 1$. Accordingly, evaluation metrics used in Multi-label classification should be generalization of techniques and measures used in Multi-class classification. It is easy to see that by imposing the same limitation, proposed algorithm for the confusion matrix computation reduces to a simpler form which is used to compute confusion matrix for Multi-class classification problem.

The proposed technique emerged from our work on a specific multi-label classification problem. In our problem, classes are loosely defined and often share common features. Our intention was not to propose a new evaluation metrics, but to develop a tool to gain deeper understanding of data set and classifier operation. Therefore, the proposed technique is indeed used in decision making in a specific case study. Specific details from our work are omitted and would exceed the scope of the paper. Intention of this paper is to provide an insight to this technique to the classifier developers community and expect feedback on specific issues of the technique.

The proposed technique emerged from our work on multilabel classifier for scene understanding which will be used as a case study for this method. We expect to propose even more general solution without some minor limitations mentioned in this method description. In the future we also expect to deliver tools for automatic extraction of confusion matrix semantics as a result of comprehensive testing on several case study multi-label classifiers. In that stage we will also consider computational complexity aspect of the proposed solution. According to ours' best knowledge, at this moment there are no other proposed methods for constructing multi-label confusion matrix, and we consider this a good basis for developing a general solution to this problem, or a specialized method for evaluation of a specific problems.

ACKNOWLEDGEMENTS

This work is partly supported by the Ministry of Science, Education and Sport of the Republic of Croatia under VIF Grant through project ViO - Vision Based Intelligent Observers.

REFERENCES

[1] B. Pereira, R., Plastino, A., Zadrozny, B., Merschmann, L.: Correlation analysis of performance measures for multi-label classification. Information Processing and Management 54, 359–369 (05 2018). https://doi.org/10.1016/j.ipm.2018.01.002

[2] Braović, M., Stipaničev, D., Krstinić, D.: Cogent confabulation based expert system for segmentation and classification of natural landscape images. Advances in Electrical and Computer Engineering 17, 85–94 (2017). https://doi.org/10.4316/AECE.2017.02012, http://dx.doi.org/10.4316/AECE.2017.02012

[3] Canbek, G., Sagiroglu, S., Taskaya Temizel, T., Baykal, N.: Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In: 2017 International Conference on Computer Science and Engineering (UBMK). pp. 821–826 (10 2017). https://doi.org/10.1109/UBMK.2017.8093539

[4] Fawcett, T.: An introduction to roc analysis. Pattern Recogn. Lett. 27(8), 861–874 (Jun 2006). https://doi.org/10.1016/j.patrec.2005.10.010, http://dx.doi.org/10.1016/j.patrec.2005.10.010

[5] Galindo, E.L.G., Ventura, S.: Multi-label learning: a review of the state of the art and ongoing research. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 4, 411–444 (2014)

[6] GiraldoForero, A.F., Jaramillo-Garzón, J., Castellanos-Dominguez, G.: Evaluation of examplebased measures for multi-label classification performance. In: Lecture Notes in Computer Science. vol. 9043, pp. 557–564 (04 2015). https://doi.org/10.1007/978-3-319-16483-054

[7] Goutte, C., Gaussier, E.: A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In: Proceedings of the 27th European Conference on Advances in Information Retrieval Research. pp. 345– 359. ECIR'05, Springer-Verlag, Santiago de Compostela, Spain (2005)

[8] Haj Mohamad, T., Chen, Y., Chaudhry, Z., Nataraj, C.: Gear fault detection using recurrence quantification analysis and support vector machine. Journal of Software Engineering and Applications 11 (05 2018). https://doi.org/10.4236/jsea.2018.115012

[9] Hardin, P.J., Shumway, M.J.: Statistical signficance and normalized confusion matrices. Photogrammetric Engineering and Remote Sensing 63, 735–740 (1997)

[10] Kavitha, A.S., Shivakumara, P., Kumar, G.H., Lu, T.: Text segmentation in degraded historical document images. Egyptian Informatics Journal 17(2), 189–197 (2016). https://doi.org/https://doi.org/10.1016/j.eij.2015.11.003

[11] Koço, S., Capponi, C.: On multi-class learning through the minimization of the confusion matrix norm. Journal of Machine Learning Research 29 (03 2013)

[12] Koyejo, O., Ravikumar, P., Natarajan, N., Dhillon, I.S.: Consistent multilabel classification. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. pp. 3321–3329. NIPS'15, MIT Press, Montreal, Canada (2015)

[13] Labatut, V., Cherifi, H.: Evaluation of performance measures for classifiers comparison. Ubiquitous Computing and Communication Journal 6, 21–34 (11 2011)

[14] Li, S., Deng, W.: Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. International Journal of Computer Vision (Nov 2018). https://doi.org/10.1007/s11263-018-1131-1, https://doi.org/10.1007/s11263-018-1131-1

[15] Madjarov, G., Kocev, D., Gjorgjevikj, D., Deroski, S.: An extensive experimental comparison of methods for multi-label learning. Pattern Recogn. 45(9), 3084–3104 (Sep 2012). https://doi.org/10.1016/j.patcog.2012.03.004, http://dx.doi.org/10.1016/j.patcog.2012.03.004

[16] Miller, G.A., Nicely, P.E.: An analysis of perceptual confusions among some english consonants. The Journal of the Acoustical Society of America 27, 338–352 (03 1955). https://doi.org/10.1121/1.1907526

[17] Tidake, V.S., Sane, S.S.: Multi-label classification: a survey. International Journal of Engineering and Technology 7(4.19), 1045–1054 (2018). https://doi.org/10.14419/ijet.v7i4.19.28284, https://www.sciencepubco.com/index.php/ijet/article/

[18] Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM) 3(3), 1–13 (2007), https://EconPapers.repec.org/RePEc:igg:jdwm00:v:

[19] Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: Cnn-rnn: A unified framework for multi-label image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2285–2294. Las Vegas, Nevada, USA (2016)

[20] Weisstein, E.W.: Iverson Bracket From Math- World – A Wolfram Web Resource (2019), http://mathworld.wolfram.com/IversonBracket.html, [Online; accessed 9-March-2019]

[21] Zhang, M.L., Cheng Zhou, Z.: A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering 26(8):1819-1837

AUTHORS

Damir Krstinić received the Ph.D degree from the University of Split, Split, Croatia in 2008. He is Associate Professor on Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture of the University of Split, Department for Modelling and Intelligent Systems. His main area of research interest are computer vision, image understanding and machine learning. He has been involved in several research and technological projects related to forest fire research and is project leader of Wildfire Early Detection and Monitoring System for Croatian Forests.

Maja Braović received her Bachelor's and Master's degrees in Computer Science from the University of Split, Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, in 2008 and 2010, respectively. She received her PhD degree in Artificial Intelligence from the same Faculty in 2015, where she is also currently a postdoctoral researcher. Her main areas of research interest include artificial intelligence, computer vision, image understanding and natural language processing.

Ljiljana Šerić is associate professor at University of Split, Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture. She received her PhD in electrical engineering and computer science in 2010. She is a member of Department for Modelling and Intelligent Systems and Centre for wildfire research. She has co-authored more than 50 scientific papers. Her research interests are focused on artificial intelligence and web and distributed information systems.

Dunja Božić - Štulić received her Master of Engineering degree in Electronics and Computer engineering in 2014, from Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, University of Split, Croatia. She is currently working as research assistant at Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, University of Split, Croatia and is PhD candidate. Dunja is co-author of 5 conference paper's and 2 scientific paper's. (http://bib.irb.hr/lista-radova?autor=355870). Her research interests include machine learning, deep learning and artificial intelligence.