Information-theoretic Measures for Objective Evaluation of Classifications

HU Bao-Gang1, 2, HE Ran1, YUAN Xiao-Tong2

Abstract This work presents a systematic study of objective evaluations of abstaining classifications using information-theoretic measures (ITMs). First, we define objective measures as the ones which do not depend on any free parameter. According to this definition, technical simplicity for examining "objectivity" or "subjectivity" is directly provided for classification evaluations. Second, we propose 24 normalized ITMs for investigation, which are derived from either mutual information, divergence, or cross-entropy. Contrary to conventional performance measures that apply empirical formulas based on users' intuitions or preferences, the ITMs are theoretically more general for realizing objective evaluations of classifications. They are able to distinguish "error types" and "reject types" in binary classifications without the need to inputting data of cost terms. Third, to better understand and select the ITMs, we suggest three desirable features for classification assessment measures, which appear more crucial and appealing from the viewpoint of classification applications. Using these features as "meta-measures", we can reveal the advantages and limitations of ITMs from a higher level of evaluation knowledge. Numerical examples are given to demonstrate our claims and compare the differences among the proposed measures. The best measure is selected in terms of the meta-measures, and its specific properties regarding error types and reject types are analytically derived.

Key words Abstaining classifications, error types, reject types, entropy, similarity, objectivity, information-theoretic measures (ITMs)


DOI 10.1016/S1874-1029(11)60289-9

The selection of evaluation measures for classifications has received increasing attentions from researchers on various application fields[1−7]. It is well known that evaluation measures, or criteria, have a substantial impact on the quality of classification performance. The problem of how to select evaluation measures for the overall quality of classifications is difficult, and there appears no universal answer to this. Due to various types of evaluation measures have been used in classification applications. Taking binary classification as an example, more than 30 metrics have been applied for assessing the quality of classifications and their algorithms as given in Table 1 of [5]. Most of the metrics listed in this table can be considered as a type of performance-based measures. In practice, other types of evaluation measures, such as information-theoretic measures (ITMs), have also commonly been used in machine learning[8−9]. The typical information-based measure used in classifications is the cross entropy[10]. In a recent work[11], Hu et al. derived an analytical formula of the Shannon-based mutual information measure with respect to a confusion matrix. Significant benefits were observed from the measure, such as its generality even for cases of classifications with a reject option, and its objectivity in naturally balancing performance-based measures that may conflict with one another (such as precision and recall). The objectivity was achieved from the perspective that an information-based measure does not require knowledge of cost terms in evaluating classifications. This advantage is particularly important in studies of abstaining classifications[12−14] and cost sensitive learning[15−18].

where cost terms may be required as input data for evaluations. Generally, if no cost terms are assigned to evaluations, it implies that the zero-one cost functions are applied[19]. In such situations, classification evaluations without a reject option may still be applicable and useful in class-balanced datasets. Problematic, or unreasonable, results will be obtained for evaluations in situations where classes are highly skewed in the datasets[20] if no specific cost terms are given.

In this work, for simplifying discussions, we distinguish, or decouple, two study goals in evaluation studies, namely, evaluation of classifiers and evaluation of classifications. The former goal concerns more about evaluation of algorithms in which classifiers are applied. From this evaluation, designers or users can select the best classifier. The latter goal is to evaluate classification results without concerning which classifier is applied. This evaluation aims more on result comparisons or measure comparisons. One typical example was demonstrated by Mackay[20] for highlighting the difficulty in classification evaluations. He showed two specific confusion matrices, $C_D$ and $C_E$, in binary classifications with a reject option:

$$C_D = \begin{bmatrix} 74 & 6 & 10 \\ 0 & 9 & 1\end{bmatrix}, \quad C_E = \begin{bmatrix} 78 & 6 & 6 \\ 0 & 5 & 5\end{bmatrix},$$

where $C = \begin{bmatrix} TN & FP & RN \\ FN & TP & RP\end{bmatrix}$

(1)

1. Chinese-French Joint Laboratory for Computer Science, Control and Applied Mathematics, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100090, P. R. China. 2. Graduate University of Chinese Academy of Sciences, Beijing 100090, P. R. China. 3. Department of Statistics, Rutgers University, New Jersey 08816, USA
types without using cost terms? level knowledge for abstaining classification evaluations?

citations, “ent losses in applications. For example, in medical appli-
cations, (referring to Table 3 of [11]) observed that

contribution of the work is derived from the following three

1) We define the “proper” features, also called “meta-
measures”, for selecting candidate measures in the context of
abstaining classification evaluations. These features will
assist users in understanding advantages and limitations of
evaluation measures from a higher level of knowledge.

2) We examine most of the existing information measures
in a systematic investigation of “error types” and “reject
types” for objective evaluations. We hope that the more
than 20 measures investigated are able to enrich the cur-
rent bank of classification evaluation measures. For the
best measure in terms of the meta-measures, we present a
theoretical confirmation of its desirable properties regard-
ing error types and reject types.

3) We reveal the intrinsic shortcomings of information
measures in evaluations. The discussions are intended to
be applicable to a wider range of classification problems,
such as similarity ranking. The finding of the shortcomings
of information measures is so important that we are able to
employ the measures reasonably in interpreting evaluation
results.

To address classification evaluations with a reject option,
we assume that the only basic data available for classifica-
tion evaluations is a confusion matrix, without inputting
data of cost terms. The rest of this paper is organized as
follows. In Section 1, we present related work for the se-
lection of evaluation measures. For seeking “proper” mea-
sures, we propose several desirable features in the context of
classifications in Section 2. Three groups of normalized
information measures are proposed along with their intrin-
sic shortcomings in Sections 3–5, respectively. Several
numerical examples, together with discussions, are given in
Section 6. Finally, in Section 7 we conclude the work.

1 Related work

In classification evaluations, a measure based on clas-
sification accuracy has traditionally been used with some
success in numerous cases[19]. This measure, however, may
suffer serious problems in reaching intuitively reasonable
results from certain special cases of real-world classifica-
tion problems[35]. The main reason for this is that a single
measure of accuracy does not take error types into account.

To overcome the problems of accuracy measures, re-
searchers have developed many sophisticated approaches
for classification assessment[21–25]. Among these, two
commonly-used approaches are receiver operating char-
acteristic (ROC) curves and area under curve (AUC)
measures[1, 26]. ROC curves provide users with a very fast
evaluation approach via visual inspections, but this is only
applicable in limited cases with specific curve forms (for
example, when one curve is completely above the other).
AUC measures are more generic for ranking classifications
without constraints on curve forms. In a study of binary
classifications, a formal proof was given by Ling et al.[31],
showing that AUC is a better measure than accuracy from
the definitions of both statistical consistency and discrimi-
nancy. Sophisticated AUC measures were reported recently

Drummond et al.[27] proposed a visualization technique
called “cost curve”, which is able to take cost terms into ac-
count for showing confidence intervals on classifier’s perform-
cance. Japkowicz[3] presented convincing examples show-
ing the shortcomings of the existing evaluation methods, in-
cluding accuracy, precision vs. recall, and ROC techniques.
The findings from the examples further confirmed the need
for methods using measure-based functions[28]. The main
idea behind measure-based functions is to form a single
function with respect to a weighted summation of multiple
measures. The measure function is able to balance a trade-
off among the conflicting measures, such as precision and
recall. However, the main difficulty arises in the selection
of balancing weights for the measures[5]. In most cases,
users rely on their preferences and experiences in assigning
the weights, which imposes a strong degree of subjectivity
on the evaluation results.

Classification evaluations become more complicated if a
classifier abstains from making a prediction when the out-
come is considered unreliable for a specific sample. In
this case, an extra class, known as the “reject” or “un-
known” class, is added to the classification. In recent
years, the study of abstaining classifiers has received much
attention[14, 12–14, 29–31]. With complete data of a full cost
matrix, they were able to assess the classifications. If one
term of the cost matrix was missing, such as on a reject cost
term, the approaches for classification evaluations generally
failed. Moreover, because in most situations the cost terms
are given by users, this approach is basically a subjective
evaluation in applications. Vanderlooy et al.[32] further
investigated the ROC isometrics approach which does not
rely on information from a cost matrix. This approach,
however, is only applicable to binary classification prob-
lems.

A promising study of objective evaluations of classifica-
tions is attributed to the introduction of information the-
ory. Kvalseth[33] and Wickens[34] derived normalized mu-
tual information (NMI) measures in relation to a contingency table. Further pioneering studies on the classification problems were conducted by Finn[35] and Forbes[36]. Forbes discussed the problem that NMI does not share monotonic property with the other performance measures, such as accuracy or $F$-measure. Several different definitions for information measures have been reported in studies of classification assessment, such as information scores[37] and KL divergence[38]. Accordingly, Yao et al.[39] and Tan et al. summarized many useful information measures for studies of associations and attribute importance. Significant efforts were made on discussing the desired properties of evaluation measures[39]. Principe et al.[9] proposed a framework of information theoretic learning (ITL) that included most types of learnings, such as classifications. Within this framework, the learning criteria were the mutual information defined from the Shannon and Renyi entropies. Two quadratic divergences, namely, the Euclidean and Cauchy-Schwartz distances were also proposed.

From the comparison perspective between ITL[9] and conventional performance, Wang et al.[40] derived for the first time the nonlinear relations between mutual information and the conventional performance measures (accuracy, recall, and precision) for binary classification problems. They extended the investigation into abstaining classification evaluations for multiple classes[11]. Their method was based solely on the confusion matrix. For gaining the theoretical properties, they derived the extremum theorems concerning mutual information measures. One of the important findings from the local minimum theorem is the theoretical revelation of the non-monotonic property of mutual information measures with respect to the diagonal terms of a confusion matrix. This property may cause irrational evaluation results from some data in classifications. They confirmed this problem by examining specific numerical examples. Theoretical investigations are still missed for other information measures, such as divergence-based and cross-entropy based ones.

2 Objective evaluations and meta-measures

This work focuses on objective evaluations of classifications. While Berger[41] stressed four points from a philosophical perspective for supporting objective Bayesian analysis, it seems that few studies in the literature address the "objectivity" issue in the study of classification evaluations. Some researchers[39] may call their measures to be objective ones without defining them formally. Considering that "objectivity" is a more philosophical concept without a well accepted definition, we propose a scheme for defining "objective evaluations" from the viewpoints of practical implementation and examination.

Definition 1 (Objective evaluations and measures). An objective evaluation is an assessment expressed by a function that does not contain any free parameter. This function is called an objective measure.

Remark 1. When a free parameter is used to define a measure, it usually carries a certain degree of subjectivity in evaluations. Therefore, according to this definition, a measure based on cost terms[40] as free parameters does not lead to an objective evaluation. Definition 1 may be conservative, but nevertheless, provides technical simplicity for examining "objectivity" or "subjectivity" directly with respect to the existence of free parameters. In some situations, Definition 1 can be relaxed by including free parameters, but they all have to be determined solely from the given dataset.

Definition 2 (Datasets in classification evaluations with a reject option). A reject option is sometimes considered for classifications in which one may assign samples to a reject or unknown class. Evaluations of classification with a reject option apply to two datasets, namely, the output (or prediction) dataset \( \{y_k\}_{k=1}^{n+m} \), which is a realization of discrete random variable $Y$ valued on set \( \{1,2,\cdots,m+1\} \); and the target dataset \( \{t_k\}_{k=1}^{n} \) valued on set \( \{1,2,\cdots,m\} \); where $n$ is the total number of samples, and $m$ is the total number of classes. A sample identified as a reject class is represented by $y_k = m+1$.

Remark 2. The term "abstaining classifiers" has been widely used in classification problems with a reject option[4, 12]. However, most studies of abstaining classifications required cost matrices for their evaluations. The definition given above exhibits more generic scenarios in classification evaluations, because it does not require knowledge of cost terms for error types and reject types.

Definition 3 (Augmented confusion matrix and its constraints[11]). An augmented confusion matrix includes one column for the reject class, which is added to a conventional confusion matrix:

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} & c_{1(m+1)} \\ c_{21} & c_{22} & \cdots & c_{2m} & c_{2(m+1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mm} & c_{m(m+1)} \end{bmatrix}$$

where $c_{ij}$ represents the sample number of the $i$-th class that is classified as the $j$-th class. The row data corresponds to the actual classes, while the column data corresponds to the predicted classes. The last column represents the reject class. The relations and constraints of an augmented confusion matrix are

$$C_i = \sum_{j=1}^{m+1} c_{ij}, \quad C_i > 0, c_{ij} \geq 0, i = 1,2,\cdots,m$$

where $C_i$ is the total number of the $i$-th class, which is generally known in classification problems.

Definition 4 (Error types and reject types). Following the conventions in binary classifications[42], we denote $c_{12}$ and $c_{21}$ as "type I error" and "type II error" respectively; $c_{13}$ and $c_{23}$ as "type I reject" and "type II reject" respectively.

Definition 5 (Normalized information measure). A normalized information measure, denoted as $NI(T,Y) \in [0,1]$, is a function based on information theory, which represents the degree of similarity between two random variables $T$ and $Y$.

In principle, we hope that all $NI$ measures satisfy three important properties or axioms of metrics[19, 43], supposing $Z$ is another random variable:

1) $NI(T,Y) = 1$, if and only if $T = Y$ (the identity axiom);
2) $NI(T,Y) + NI(Y,Z) \geq NI(T,Z)$ (the triangle inequality);
3) $NI(T,Y) = NI(Y,T)$ (the symmetry axiom).

Remark 3. Violations of properties of metrics are possible in reaching reasonable evaluations of classifications. For example, the triangle inequality and symmetry properties can be relaxed without changing the ranking orders among classifications if their evaluation measures are applied con-
sistently. However, the identity property is indicated only for the relation $T = Y$ (assuming $T$ is padded with a zero-value term to make it be of the same size as $Y$), and does not guarantee an exact solution ($t_k = y_k$) in classifications (see Theorems 1 and 4 given later). If a violation of metric properties occurs, the $NIs$ are referred to as measures, rather than metrics.

For classification evaluations, we consider the generic properties of metrics not to be as crucial in comparisons as certain specific features. In this work, we focus on specific features that, though not mathematically fundamental, are more necessary in classification applications. To select “better” measures for objective evaluations of classifications, we propose the following three desirable features together with their heuristic reasons.

**Feature 1 (Monotonicity with respect to the diagonal terms of the confusion matrix).** The diagonal terms of the confusion matrix represent the exact classification numbers for all the samples. Or they reflect the coincident numbers between $t$ and $y$ from a similarity viewpoint. When one of these terms changes, the evaluation measure should change in a monotonous way. Otherwise, any non-monotonic measure may fail to provide a rational result for ranking classifications correctly. This feature is originally proposed for describing the strength of agreement (or similarity) if the matrix is a contingency table.[39]

**Feature 2 (Variation with reject rate).** To improve classification performance, a reject option is often used in engineering applications.[12] Therefore, we suggest that a measure should be a scalar function for both classification accuracy and reject rates. Such a measure could be evaluated based solely on a given confusion matrix from a single operating point in the classification. This is different from the AUC measures that are calculated based on an “error-reject” curve,[20,31] from multiple operating points.

**Feature 3 (Intuitively consistent costs among error types and reject types).** This feature is derived from the principle of our conventional intuitions when dealing with error types in classifications. It is also extended to reject types. Two specific intuitions are adopted for binary classifications. First, a misclassification or rejection from a small class will cause a greater cost than that from a large class. This intuition represents a property called “within error types and reject types”. Second, a misclassification will produce a greater cost than a rejection from the same class, which is called “between error and reject types” property. If a measure is able to satisfy the intuitions, we refer to its associated costs as being “intuitively consistent”. Exceptions may exist to the intuitions above, but we consider them as a very special case.

At this stage, it is worth discussing on “objectivity” in evaluations because one may doubt correctness of the intuitions above and the terms “desirable” or “intuitions” in a study of objective evaluations. The three features seem to be “problematic” in terms of providing a general concept of “objectivity”, because no human bias should be applied in the objective judgment of evaluation results. The following discussions justify the proposal of requiring desirable or proper features for objective measures. On one hand, we recognize that any evaluation will imply a certain degree of “subjectivity”, since evaluations exist only as a result of human judgment. For example, every selection of evaluation measures, even of objective ones, will rely on possible sources of “subjectivity” from users. On the other hand, engineering applications do concern about objective evaluations.[36,39]

However, to the authors’ best knowledge, a technical definition, or criterion, seems missing for determining objective or subjective measures in evaluations of classifications. For overcoming possible confusion and vagueness, we set Definition 1 as a practical criterion for examining whether a classification evaluation holds “objectivity” or not. If a measure satisfies this definition, it will always retain the property of “objective consistency” in evaluating the given classification results. The three “desirable” features, though based on “intuitions” with “subjectivity”, do not destroy the criterion of “objectivity” in classification evaluations. Therefore, it is logically correct to discuss “desirable” features of objective measures as long as the measures satisfy Definition 1 for keeping the defined “objectivity”.

Note that all desirable features above are derived from our intuitions on general cases of classification evaluations. Other items may be derived for a wider examination of features. For example, Forbes[36] proposed six “constraints on proper comparative measures”, namely, “statistically principled, readily interpretable, generalizable to k-class situations, not different to the special status, reflective of agreement, and insensitive to the segmentation”. However, we consider the three features proposed in this work to be more crucial, especially as Feature 3 has never been concerned in previous studies of classification evaluations. Although Features 2 and 3 may share a similar meaning, they are presented individually to highlight their specific concerns.

We can also call the desirable features “meta-measures”, since these are defined to be qualitative and high-level “measures about measures”. In this work, we apply meta-measures in our investigation of information measures. The examination with respect to the meta-measures enables clarification of the causes of performance differences among the examined measures in classification evaluations. It will be helpful for users to understand advantages and limitations of different measures, either objective- or subjective-ones, from a higher level of evaluation knowledge.

3 Normalized information measures based on mutual information

All $NI$ measures applied in this work are divided into one of three groups, namely, mutual-information based, divergence based, and cross-entropy based groups. In this section, we focus on the first group. Each measure in this group is derived directly from mutual information for representing the degree of similarity between two random variables. For the purpose of objective evaluations, as suggested by Definition 1 in the previous section, we eliminate all candidate measures defined from the Renyi or Jensen entropies[39,44] since they involve a free parameter. Therefore, without adding free parameters, we only apply the Shannon entropy to information measures[45]:

$$H(Y) = -\sum_y p(y) \log_2 p(y) \quad (4)$$

where $Y$ is a discrete random variable with probability mass function $p(y)$. Then, mutual information is defined as[45]:

$$I(T,Y) = \sum_y \sum_t p(t,y) \log_2 \frac{p(t,y)}{p(t)p(y)} \quad (5)$$

where $p(t,y)$ is the joint distribution for the two discrete random variables $T$ and $Y$, and $p(t)$ and $p(y)$ are called marginal distributions that can be derived from

$$p(t) = \sum_y p(t,y), \quad p(y) = \sum_t p(t,y) \quad (6)$$
Sometimes, the simplified notations for \( p_{ij} = p(t, y) = p(t = t_i, y = y_j) \) are used in this work. Table 1 lists the possible normalized information measures within the mutual-information based group. Basically, they all make use of (5) in their calculations. The main differences are due to the normalization schemes. In applying the formulas for calculating \( NI_k \), one generally does not have an exact \( p(t, y) \). For this reason, we adopt an empirical joint distribution defined below for the calculations.

Definition 6 (Empirical joint distribution and empirical marginal distributions). An empirical joint distribution is defined from the frequency means for the classifications. The subscript "empirical" is given for denoting the empirical terms. The empirical marginal distributions are

\[
P_e(t) = \frac{\sum_i c_{ti}}{n}, \quad i = 1, 2, \ldots, m
\]

\[
P_e(y) = \frac{1}{n} \sum_{i=1}^{m} c_{ij}, \quad j = 1, 2, \ldots, m + 1
\]

Definition 7 (Empirical mutual information). The empirical mutual information is given by

\[
I_e(T, Y) = \sum_t \sum_y P_e(t, y) \log_2 \frac{P_e(t, y)}{P_e(t)P_e(y)} = \\
\sum_{i=1}^{m} \sum_{j=1}^{m+1} \frac{c_{ij}}{n} \left( \log_2 \frac{c_{ij}}{C_i \sum_{i=1}^{m+1} c_{ij}/n} \right) \text{sgn}(c_{ij})
\]

where \( \text{sgn}(\cdot) \) is a sign function for satisfying the definition of \( H(0) = 0 \). For the sake of simplicity of expressions, we hereafter neglect the sign function.

Definitions 6 and 7 provide users with a direct means for applying information measures through the given data of the confusion matrix. For the sake of simplicity, we adopt the empirical distributions, or \( p_{ij} \) \( \approx \) \( P_{ij} \), for calculating all \( NI_k \) and deriving the theorems, but removing their associated subscript "empirical". Note that the notation of \( NI_k \) in Table 1 differs from the others for calculating mutual information, where \( I_M(T, Y) \) is defined as “modified mutual information”. The calculation of \( I_M(T, Y) \) is carried out based on the intersection of \( T \) and \( Y \). Hence, when using (8), the intersection requires that \( I_M(T, Y) \) incorporate a summation of \( j \) over 1 to \( m \), instead of \( m + 1 \). This definition is beyond mathematical rigor, but \( NI_k \) has the same properties of metrics as \( NI_1 \). It was originally proposed to overcome the problem of unchanging values in \( NI_k \) if rejections are made within only one class (refer to Table 3 of [11]). The following three theorems are derived for all \( NI_k \) in this group.

Theorem 1. Within all \( NI_k \) measures in Table 1, when \( NI(T, Y) = 1 \), the classification without a reject class may correspond to the case of either an exact classification \( (y_k = t_k) \) or a specific misclassification \( (y_k \neq t_k) \). The specific misclassification can be fully removed by simply exchanging labels in the confusion matrix (referring to Table 4 of [11]).

Proof. If \( NI(T, Y) = 1 \), we can obtain the following conditions from (8) for classifications without a reject class:

\[
p_{ij} = p(t = t_i) \approx P_e(t = t_i) = C_i/n, \quad p_{ij} = 0,
\]

\( i, j, k = 1, 2, \ldots, m, \quad k \neq i \)

These conditions describe the specific confusion matrix where only one non-zero term appears in each column (with the exception of the last \( (m+1)\)-th column). When \( j = i \), \( C \) is a diagonal matrix for representing an exact classification \( y_k = t_k \). Otherwise, a specific misclassification exists for which a diagonal matrix can be obtained by exchanging labels in the confusion matrix (referring to Table 4 of [11]).

Remark 4. Theorem 1 describes that \( NI(T, Y) = 1 \) presents a necessary, but not sufficient, condition of an exact classification.

Theorem 2. For abstaining classification problems, when \( NI(T, Y) = 0 \), the classifier generally reflects a misclassification. One special case is that all samples are considered to be one of \( m \) classes, or be a reject class.

Proof. For \( NI_k \) defined in Table 1, \( NI(T, Y) = 0 \), if and only if \( I(T, Y) = 0 \). According to information theory, the following conditions hold based on the given marginal distributions (or the empirical ones if a confusion matrix is used):

\[
I(T, Y) = 0, \quad \text{if and only if} \quad p(t, y) \approx p(t)p(y)
\]

The conditional part in (10) can be rewritten as \( p_{ij} = p(t = t_i)p(y = y_j) \). From the constraints in (3), \( p(t = t_i) > 0 \) \( (i = 1, 2, \ldots, m) \) can be obtained. For classification solutions, there should exist at least one term for \( p(y = y_j) > 0 \) \( (j = 1, 2, \ldots, m + 1) \). Therefore, at least one non-zero term

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Formula on ( NI_k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( NI_1(T, Y) )</td>
<td>( I(T, Y) )</td>
</tr>
<tr>
<td>2</td>
<td>( NI_2(T, Y) )</td>
<td>( I(T, Y) )</td>
</tr>
<tr>
<td>3</td>
<td>( NI_3(T, Y) )</td>
<td>( I(T, Y) )</td>
</tr>
<tr>
<td>4</td>
<td>( NI_4(T, Y) )</td>
<td>( I(T, Y) )</td>
</tr>
<tr>
<td>5</td>
<td>( NI_5(T, Y) )</td>
<td>( I(T, Y) )</td>
</tr>
<tr>
<td>6</td>
<td>( NI_6(T, Y) )</td>
<td>( I(T, Y) )</td>
</tr>
<tr>
<td>7</td>
<td>( NI_7(T, Y) )</td>
<td>( I(T, Y) )</td>
</tr>
<tr>
<td>8</td>
<td>( NI_8(T, Y) )</td>
<td>( I(T, Y) )</td>
</tr>
<tr>
<td>9</td>
<td>( NI_9(T, Y) )</td>
<td>( I(T, Y) )</td>
</tr>
</tbody>
</table>

Table 1. \( NI \) measures within the mutual-information based group
for \( p_{ij} > 0 \) \((i \neq j)\) must be obtained. This non-zero term corresponds to the off-diagonal term in the confusion matrix, which indicates that a misclassification has occurred. When all samples have been identified as one of the classes (referring to \( M2 \) in Table 4 of [11]), \( NI = 0 \) should be obtained.

**Remark 5.** Equation (10) gives the statistical reason for zero mutual information, that is, the two random variables are “statistically independent”. Theorem 2 demonstrates an intrinsic reason for local minima in \( NI_s \).

**Theorem 3.** The \( NI \) measures defined by the Shannon entropy generally do not exhibit a monotonic property with respect to the diagonal terms of a confusion matrix.

**Proof.** Based on [11], we arrive at simpler conditions for the local minima about \( I(T,Y) \) for the given confusion matrix:

\[
C = \begin{bmatrix}
\cdots & 0 & 0 & \cdots \\
0 & c_{i,i} & 0 & 0 \\
0 & 0 & c_{i,i+1} & 0 \\
\cdots & 0 & 0 & \cdots \\
\end{bmatrix}
\]

if \( c_{i,i} = c_{i,i+1} = c_{i+1,i} = c_{i+1,i+1} = 0 \) \((i \neq j)\) \((\cdot) \)

(11)

The local minima are obtained because the four given non-zero terms in (11) produce zero (or the minimum) contribution to \( I(T,Y) \). Suppose a generic form is given for \( N(I(T,Y)) = g(I(T,Y)) \), where \( g(\cdot) \) is a normalization function. From the chain rule of derivations, it can be seen that the conditions do not change for reaching the local minima.

**Remark 6.** The non-monotonic property of the information measures implies that these measures may suffer from an intrinsic problem of local minima for classification rankings (referring to \( M19 \sim M20 \) in Table 4 of [11]). Or according to Feature 1 of the meta-measures, a rational result for the classification evaluations may not be obtained due to the non-monotonic property of the measures. This shortcoming has not been theoretically derived in previous studies[35–36, 39].

### 4 Normalized information measures based on information divergence

In this section, we propose normalized information measures based on the definition of information divergence. In Table 2, we summarize the commonly-used divergence measures, which are denoted as \( D_k(T,Y) \) and represents dissimilarity between the two random variables \( T \) and \( Y \). In this and next sections, we apply the following notations for defining marginal distributions:

\[
p_y(z) = p_{y}(t = z) = p(t) \quad \text{or} \quad p_y(y = z) = p(y) \quad (12)
\]

where \( z \) is a possible scalar value that \( t \) or \( y \) can take. For a consistent comparison with the previous normalized information measures, we adopt the following transformation on \( D_k^{[38]} \):

\[
NI_k = e^{-D_k} \quad (13)
\]

This transformation provides both inverse and normalization functionalities. It does not introduce any extra parameters, and presents a high degree of simplicity as in derivation for examining the local minima. Two more theorems are derived by following a similar analysis as in the previous section.

**Theorem 4.** For all \( NI \) measures in Table 2, when \( N(I(T,Y)) = 1 \), the classifier corresponds to the case of either an exact classification, or a specific misclassification. Generally, the misclassification in the latter case cannot be removed by switching labels in the confusion matrix.

**Proof.** When \( p_y(z) = p_t(z) \), it is always the case that \( N(I(T,Y)) = 1 \). However, general conditions can be given for \( p_y(z) = p_t(z) \) as follows:

\[
p_y(y = z) = p_t(t = z) \quad \text{or} \quad \sum_j p_{ji} = \sum_j p_{ij}, \quad i = 1, 2, \ldots, m \quad (14)
\]

Equation (14) implies two cases of classifications for \( D_k(T,Y) = 0 \) \((k = 10, \ldots, 20)\). One of these corresponds to an exact classification \((y_k = t_k)\), while the other is the result of a specific misclassification that shows the relationship of \( y_k \neq t_k \). For the case \( p_y(z) = p_t(z) \), switching of labels in the confusion matrix to remove misclassification generally destroys the relation for \( p_y(z) = p_t(z) \) at the same time. Considering the relation as a necessary condition for a perfect classification, the misclassification cannot be removed through a label switching operation.

### Table 2 Information measures within the divergence based group

<table>
<thead>
<tr>
<th>No.</th>
<th>Name of ( D_k )</th>
<th>Formula on ( D_k ) ((NI_k = e^{-D_k}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>ED-Quadratic divergence[^{[9]}]</td>
<td>( D_{10} = QD_{ED}(T,Y) = \sum \left( p(z) - p_t(z) \right)^2 )</td>
</tr>
<tr>
<td>11</td>
<td>CS-Quadratic divergence[^{[9]}]</td>
<td>( D_{11} = QD_{CS}(T,Y) = \log_{2} \sum \left( \frac{p_t(z)}{2} \right)^2 p(y) )</td>
</tr>
<tr>
<td>12</td>
<td>KL divergence[^{[48]}]</td>
<td>( D_{12} = KL(T,Y) = \sum p_t(z) \log \frac{p_t(z)}{p_t(y)} )</td>
</tr>
<tr>
<td>13</td>
<td>Bhattacharyya distance[^{[49]}]</td>
<td>( D_{13} = D_B(T,Y) = -\log_{2} \sum \left( \sqrt{p_t(z) p_y(z)} \right)^2 )</td>
</tr>
<tr>
<td>14</td>
<td>(\chi^2) (Pearson) divergence[^{[50]}]</td>
<td>( D_{14} = \chi^2(T,Y) = \sum \frac{(p_t(z) - p_y(z))^2}{p_y(z)} )</td>
</tr>
<tr>
<td>15</td>
<td>Hellinger distance[^{[50]}]</td>
<td>( D_{15} = H^2(T,Y) = \sum \left( \sqrt{p_t(z)} - \sqrt{p_y(z)} \right)^2 )</td>
</tr>
<tr>
<td>16</td>
<td>Variation distance[^{[50]}]</td>
<td>( D_{16} = V(T,Y) = \sum \left( p_t(z) - p_y(z) \right)^2 )</td>
</tr>
<tr>
<td>17</td>
<td>J divergence[^{[51]}]</td>
<td>( D_{17} = J(T,Y) = \sum p_z \log \frac{p_t(z)}{p_z} + \sum p_y(z) \log \frac{p_y(z)}{p_z} )</td>
</tr>
<tr>
<td>18</td>
<td>L (or JS) divergence[^{[51]}]</td>
<td>( D_{18} = L(T,Y) = KL(T,M) + KL(Y,M) = \sum p_t(z) p_y(z) )</td>
</tr>
<tr>
<td>19</td>
<td>Symmetric (\chi^2) divergence[^{[52]}]</td>
<td>( D_{19} = \chi^2(T,Y) = \sum \frac{(p_t(z) - p_y(z))^2}{p_y(z)} + \sum \frac{(p_y(z) - p_t(z))^2}{p_t(z)} )</td>
</tr>
<tr>
<td>20</td>
<td>Resistor average distance[^{[49]}]</td>
<td>( D_{20} = D_{RA}(T,Y) = \frac{KL(T,Y) + KL(Y,T)}{2} )</td>
</tr>
</tbody>
</table>
Remark 7. Theorem 4 suggests the caution should be applied in explaining the classification evaluations when \( N(I(T, Y) = 1). \) The maximum of the \( NIs \) from the information divergence measures only indicates the equivalence between the marginal probabilities, \( p(y) = p_1(z) \), but this is not always true for representing exact classifications (or \( y_k = t_k \)). Theorem 4 reveals an intrinsic problem when using an \( NI \) as a measure for similarity evaluations between two datasets, such as in image registration.

Theorem 5. The \( NI \) measures based on information divergence generally do not exhibit a monotonic property with respect to the diagonal terms of confusion matrix.

Proof. The theorem can be proved by examining the existence of multiple maxima for \( D_k \) or \( \sum_d \) as an example. The local minima of cross-entropy. Here, we use a binary classification as an example. The local minima of \( D_k \) or \( \sum_d \) are obtained when the following conditions exist for a confusion matrix:

\[
C = \begin{bmatrix}
C_1 - d_1 & d_1 & 0 \\
d_2 & C_2 - d_2 & 0 \\
0 & 0 & d_1 = d_2
\end{bmatrix}
\]

where \( d_1 \) and \( d_2 \) are positive integers for misclassified samples. The confusion matrix in (15) produces zero divergence \( D_k \) and therefore, \( N(I_k) = 1. \) However, changing from \( d_1 = d_2 \) always results in \( N(I_k) < 1. \) The above problem can be extended to the general classifications in (2).

Remark 8. Theorem 5 indicates another shortcoming of \( NIs \) in the information divergence group from the viewpoint of monotonicity. The reason is once again attributed to the usage of marginal distributions in calculations of divergence. The shortcoming has not been reported in previous investigations [38, 43].

5 Normalized information measures based on cross-entropy

In this section, we propose normalized information measures based on cross-entropy, which is defined for discrete random variables as

\[
H(T; Y) = - \sum_z p(z) \log_2 p(z)
\]

or \( H(Y; T) = - \sum_z p(y) \log_2 p(y) \) (16)

Note that \( H(T; Y) \) differs from joint-entropy \( H(T, Y) \) with respect to both notation and definition, and is given as [45]

\[
H(T, Y) = - \sum_{t,y} p(t, y) \log_2 p(t, y)
\]

In fact, from (16), one can derive the relation between KL divergence (see Table 2) and cross-entropy:

\[
H(T; Y) = H(T) + KL(T, Y)
\]

or \( H(Y; T) = H(Y) + KL(Y, T) \) (18)

If \( H(T) \) is considered as a constant in classification since the target dataset is generally known and fixed, we can observe from (18) that cross-entropy shares a similar meaning as KL divergence for representing dissimilarity between \( T \) and \( Y. \) From the conditions \( H \geq 0 \) and \( KL \geq 0, \) we are able to realize the normalization for cross-entropy shown in Table 3. Following similar discussions as in the previous section, we can derive that all information measures listed in Table 3 will also satisfy Theorems 4 and 5.

6 Numerical examples and discussions

This section presents several numerical examples together with associated discussions. All calculations for the numerical examples were done using the open source software Scilab [1] and a specific toolbox 2. The detailed implementation of this toolbox is described in [53]. Table 4 lists six numerical examples in binary classification problems according to the specific scenarios of their confusion matrices. We adopt the notations from [54] for the terms “correct recognition rate (CR),” “error rate (E),” and “reject rate (Rej)” and their relation:

\[
CR + E + Rej = 1
\]

In addition, we define “accuracy rate (A)” as

\[
A = \frac{CR}{CR + E}
\]

The first four classifications (or models) \( M1 \sim M4 \) are provided to show the specific differences with respect to error types and reject types. In this work, we do not concern the classifiers applied (say, neural networks or support vector machines) for evaluations, but only the resulting evaluations from any classifier. In real applications, it is common to encounter ranking classification results as in \( M1 \sim M4. \) The first two classifications of \( M1 \) and \( M2 \) share the same values for the correct recognition and accuracy rates (\( CR = A = 0.99 \)). The other two classifications, \( M3 \) and \( M4, \) have the same rates for \( Rej = 0.01, \) \( CR = 0.99, \) and \( A = 1.00. \) The data from other conventional measures, such as “precision,” “recall” and \( F1, \) are also given in Table 4. Without using extra knowledge about costs of different error types or reject types, the conventional performance measures are unable to rank the four classifications, \( M1 \sim M4, \) properly.

According to the intuitions of Feature 3, one can gain two sets of ranking orders for the four classifications \( M1 \sim M4 \) in forms of

\[
\mathbf{R}(M2) > \mathbf{R}(M1), \quad \mathbf{R}(M4) > \mathbf{R}(M3) \tag{21a}
\]

\[
\mathbf{R}(M4) > \mathbf{R}(M2), \quad \mathbf{R}(M3) > \mathbf{R}(M1) \tag{21b}
\]

where we denote \( \mathbf{R}(\cdot) \) to be a ranking operator, so that \( \mathbf{R}(M_i) > \mathbf{R}(M_j) \) expresses \( M_i \) is better than \( M_j \) in ranking. From (21), one is unable to tell the ranking order between \( M2 \) and \( M3. \) For a fast comparison, a specific letter is assigned to the ranking order of each model in Table 4 based on (21):

\[
\mathbf{R}(M4) = A, \quad \mathbf{R}(M3) = B, \quad \mathbf{R}(M2) = B, \quad \mathbf{R}(M1) = C \tag{22}
\]

The top rank “\( A \)” indicates the “best” classification (\( M4 \) in this case) of the four models. Table 4 does not distinguish ranking order between \( M2 \) and \( M3. \) However, numerical investigations using information measures will provide the ranking order from the given data. The other two models, \( M5 \) and \( M6, \) are also specifically designed for the purpose of examining information measures on Theorems 3 and 5 (or Feature 1), respectively.

Tables 5 and 6 present the results on information measures for \( M1 \sim M6, \) where the ranking orders among \( M1 \sim M4 \) are based on the calculation results of \( NIs \) with the given digits. The following discussions are achieved from the solutions to the examples.  

\[1\]http://www.scilab.org

\[2\]The toolbox is freely available as the file “confmatrix2nu.zip” at http://www.openpr.org.cn.cn.
None of the performance or information measures investigated in this work fully satisfies the meta-measures. Examining data distinguishability in $M_1 \sim M_4$, we consider the information measures from the mutual-information group to be more appropriate than those of the other groups (say, $NI_{12}$ and $NI_{22}$ do not show significant distinguishability, or value differences, for the four models).

If examining the meta-measure on Feature 3, one can observe that of all the measures in the study only $NI_2$ shows some consistency with the intuitions from the given examples (Tables 5 and 6). This result indicates that Feature 3 seems to be another difficult property for evaluation measures.

3) The results of $M_5$ and $M_6$ confirm, respectively, Theorem 3 for local minima and Theorem 5 for maxima of $NI$s. The existence of multi extrema indicates the non-monotonic property with respect to the diagonal terms of the confusion matrix, thereby exhibiting an intrinsic shortcoming of the information measures.

In general, $NI_2$ is shown to be the “best” for the given examples in evaluations. Therefore, more detailed studies,

### Table 3

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Formula on $NI_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>$NI$ based on cross-entropy</td>
<td>$NI_{k1} = \frac{H(T)}{H(T;Y)}$, $H(T;Y) = -\sum_i p_i(z) \log_2 p_i(z)$</td>
</tr>
<tr>
<td>22</td>
<td>$NI$ based on cross-entropy</td>
<td>$NI_{k2} = \frac{H(Y)}{H(Y;T)}$, $H(Y;T) = -\sum_i p_i(z) \log_2 p_i(z)$</td>
</tr>
<tr>
<td>23</td>
<td>$NI$ based on cross-entropy</td>
<td>$NI_{k3} = \frac{1}{2} \left( \frac{H(Y)}{H(Y;T)} + \frac{H(T)}{H(T;Y)} \right)$</td>
</tr>
<tr>
<td>24</td>
<td>$NI$ based on cross-entropy</td>
<td>$NI_{k4} = \frac{H(Y) + H(T)}{H(Y;T) + H(T;Y)}$</td>
</tr>
</tbody>
</table>

### Table 4

<table>
<thead>
<tr>
<th>Model</th>
<th>$M_1$</th>
<th>$M_2$</th>
<th>$M_3$</th>
<th>$M_4$</th>
<th>$M_5$</th>
<th>$M_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$(R)$</td>
<td>$(C)$</td>
<td>$(B)$</td>
<td>$(B)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
</tr>
<tr>
<td>$C$</td>
<td>90 0 0</td>
<td>89 1 0</td>
<td>90 0 0</td>
<td>89 0 1</td>
<td>57 38 0</td>
<td>89 1 0</td>
</tr>
<tr>
<td></td>
<td>1 9 0</td>
<td>0 10 0</td>
<td>0 9 1</td>
<td>0 10 0</td>
<td>3 2 0</td>
<td>1 9 0</td>
</tr>
<tr>
<td>$CR$</td>
<td>0.990</td>
<td>0.990</td>
<td>0.990</td>
<td>0.990</td>
<td>0.590</td>
<td>0.980</td>
</tr>
<tr>
<td>$Rej$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.010</td>
<td>0.010</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$Recall$</td>
<td>1.000</td>
<td>0.900</td>
<td>1.000</td>
<td>1.000</td>
<td>0.005</td>
<td>0.900</td>
</tr>
<tr>
<td>$F1$</td>
<td>0.947</td>
<td>0.952</td>
<td>0.947</td>
<td>1.000</td>
<td>0.089</td>
<td>0.900</td>
</tr>
</tbody>
</table>

### Table 5

<table>
<thead>
<tr>
<th>Model/(R)</th>
<th>$NI_1$</th>
<th>$NI_2$</th>
<th>$NI_3$</th>
<th>$NI_4$</th>
<th>$NI_5$</th>
<th>$NI_6$</th>
<th>$NI_7$</th>
<th>$NI_8$</th>
<th>$NI_9$</th>
<th>$NI_{12}$</th>
<th>$NI_{22}$</th>
<th>$NI_{13}$</th>
<th>$NI_{23}$</th>
<th>$NI_{14}$</th>
<th>$NI_{24}$</th>
<th>$NI_{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_1$</td>
<td>0.831</td>
<td>0.831</td>
<td>0.893</td>
<td>0.862</td>
<td>0.860</td>
<td>0.861</td>
<td>0.755</td>
<td>0.831</td>
<td>0.893</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>$(C)$</td>
<td>$(D)$</td>
<td>$(D)$</td>
<td>$(B)$</td>
<td>$(D)$</td>
<td>$(D)$</td>
<td>$(D)$</td>
<td>$(D)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
</tr>
<tr>
<td>$M_2$</td>
<td>0.897</td>
<td>0.897</td>
<td>0.841</td>
<td>0.869</td>
<td>0.868</td>
<td>0.869</td>
<td>0.767</td>
<td>0.841</td>
<td>0.897</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>$(B)$</td>
<td>$(C)$</td>
<td>$(C)$</td>
<td>$(C)$</td>
<td>$(C)$</td>
<td>$(C)$</td>
<td>$(C)$</td>
<td>$(C)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
</tr>
<tr>
<td>$M_3$</td>
<td>1.000</td>
<td>0.929</td>
<td>0.909</td>
<td>0.955</td>
<td>0.952</td>
<td>0.953</td>
<td>0.909</td>
<td>0.909</td>
<td>1.000</td>
<td>0.969</td>
<td>0.000</td>
<td>0.484</td>
<td>0.000</td>
<td>0.484</td>
<td>0.000</td>
<td>0.484</td>
</tr>
<tr>
<td>$(B)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(D)$</td>
<td>$(B)$</td>
<td>$(C)$</td>
<td>$(B)$</td>
<td>$(B)$</td>
<td>$(B)$</td>
<td>$(B)$</td>
<td>$(B)$</td>
</tr>
<tr>
<td>$M_4$</td>
<td>1.000</td>
<td>0.997</td>
<td>0.855</td>
<td>0.928</td>
<td>0.922</td>
<td>0.925</td>
<td>0.855</td>
<td>0.855</td>
<td>1.000</td>
<td>0.970</td>
<td>0.000</td>
<td>0.485</td>
<td>0.000</td>
<td>0.485</td>
<td>0.000</td>
<td>0.485</td>
</tr>
<tr>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(A)$</td>
<td>$(C)$</td>
<td>$(B)$</td>
<td>$(B)$</td>
<td>$(B)$</td>
<td>$(B)$</td>
<td>$(A)$</td>
<td>$(C)$</td>
<td>$(B)$</td>
<td>$(B)$</td>
<td>$(B)$</td>
<td>$(B)$</td>
<td>$(B)$</td>
<td>$(B)$</td>
<td>$(B)$</td>
</tr>
<tr>
<td>$M_5$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.374</td>
<td>0.548</td>
<td>0.461</td>
<td>0.495</td>
<td>0.461</td>
<td>0.495</td>
<td>0.461</td>
</tr>
<tr>
<td>$M_6$</td>
<td>0.731</td>
<td>0.731</td>
<td>0.731</td>
<td>0.731</td>
<td>0.731</td>
<td>0.576</td>
<td>0.731</td>
<td>0.731</td>
<td>0.731</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Table 6 Numerical examples in binary classifications (n = 100) ((R) = ranking order for the model, where R = A, B, ..., in descending order from the top)

<table>
<thead>
<tr>
<th>Model/(R)</th>
<th>N1</th>
<th>N2</th>
<th>N3</th>
<th>N4</th>
<th>N5</th>
<th>N6</th>
<th>N7</th>
<th>N8</th>
<th>N9</th>
<th>N10</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>0.9998</td>
<td>0.9998</td>
<td>0.9991</td>
<td>0.9998</td>
<td>0.9988</td>
<td>0.9997</td>
<td>0.9802</td>
<td>0.9883</td>
<td>0.9996</td>
<td>0.9977</td>
</tr>
<tr>
<td>(C)</td>
<td>(A)</td>
<td>(A)</td>
<td>(B)</td>
<td>(A)</td>
<td>(B)</td>
<td>(A)</td>
<td>(A)</td>
<td>(B)</td>
<td>(A)</td>
<td>(A)</td>
</tr>
<tr>
<td>M2</td>
<td>0.9998</td>
<td>0.9998</td>
<td>0.9992</td>
<td>0.9998</td>
<td>0.9990</td>
<td>0.9997</td>
<td>0.9802</td>
<td>0.9885</td>
<td>0.9996</td>
<td>0.9979</td>
</tr>
<tr>
<td>(B)</td>
<td>(A)</td>
<td>(A)</td>
<td>(A)</td>
<td>(A)</td>
<td>(A)</td>
<td>(A)</td>
<td>(A)</td>
<td>(A)</td>
<td>(A)</td>
<td>(A)</td>
</tr>
<tr>
<td>M3</td>
<td>0.9998</td>
<td>0.9996</td>
<td>0.9849</td>
<td>0.9926</td>
<td>0.9890</td>
<td>0.9898</td>
<td>0.9802</td>
<td>0.9897</td>
<td>0.9097</td>
<td>0.9097</td>
</tr>
<tr>
<td>(B)</td>
<td>(A)</td>
<td>(D)</td>
<td>(D)</td>
<td>(D)</td>
<td>(D)</td>
<td>(A)</td>
<td>(D)</td>
<td>(A)</td>
<td>(C)</td>
<td>(A)</td>
</tr>
<tr>
<td>M4</td>
<td>0.9998</td>
<td>0.9998</td>
<td>0.9856</td>
<td>0.9928</td>
<td>0.9899</td>
<td>0.9900</td>
<td>0.9802</td>
<td>0.9900</td>
<td>0.9000</td>
<td>0.9000</td>
</tr>
<tr>
<td>(A)</td>
<td>(A)</td>
<td>(A)</td>
<td>(C)</td>
<td>(C)</td>
<td>(C)</td>
<td>(A)</td>
<td>(C)</td>
<td>(A)</td>
<td>(C)</td>
<td>(A)</td>
</tr>
<tr>
<td>M5</td>
<td>0.7827</td>
<td>0.6473</td>
<td>0.6189</td>
<td>0.8540</td>
<td>0.6002</td>
<td>0.8129</td>
<td>0.4966</td>
<td>0.2775</td>
<td>0.7550</td>
<td>0.0455</td>
</tr>
<tr>
<td>M6</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Table 7 Numerical examples in binary classifications (n = 100) ((R) = ranking order for the model, where R = A, B, ..., in descending order from the top)

<table>
<thead>
<tr>
<th>Model</th>
<th>M1a</th>
<th>M2a</th>
<th>M3a</th>
<th>M4a</th>
<th>M1b</th>
<th>M2b</th>
<th>M3b</th>
<th>M4b</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>CR</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Rej</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>N12</td>
<td>0.756</td>
<td>0.874</td>
<td>0.876</td>
<td>0.997</td>
<td>0.720</td>
<td>0.864</td>
<td>0.849</td>
<td>0.997</td>
</tr>
<tr>
<td>(R)</td>
<td>(D)</td>
<td>(C)</td>
<td>(B)</td>
<td>(A)</td>
<td>(D)</td>
<td>(B)</td>
<td>(C)</td>
<td>(A)</td>
</tr>
</tbody>
</table>

Table 8 Classification examples in three classes (C1 = 80, C2 = 15, C3 = 5) ((R) = ranking order for the model, where R = A, B, ..., in descending order from the top)

<table>
<thead>
<tr>
<th>Model/(R)</th>
<th>M7</th>
<th>M8</th>
<th>M9</th>
<th>M10</th>
<th>M11</th>
<th>M12</th>
<th>M13</th>
<th>M14</th>
<th>M15</th>
</tr>
</thead>
<tbody>
<tr>
<td>(R)</td>
<td>(C)</td>
<td>(C)</td>
<td>(B)</td>
<td>(B)</td>
<td>(B)</td>
<td>(B)</td>
<td>(B)</td>
<td>(A)</td>
<td>(A)</td>
</tr>
<tr>
<td>C</td>
<td>80 000</td>
<td>80 000</td>
<td>80 000</td>
<td>80 000</td>
<td>80 000</td>
<td>80 000</td>
<td>80 000</td>
<td>79 010</td>
<td>79 010</td>
</tr>
<tr>
<td>1 0 4 0</td>
<td>0 1 4 0</td>
<td>0 0 4 1</td>
<td>0 0 5 0</td>
<td>0 0 5 0</td>
<td>0 0 5 0</td>
<td>0 0 5 0</td>
<td>0 0 5 0</td>
<td>0 0 5 0</td>
<td>0 0 5 0</td>
</tr>
<tr>
<td>CR</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Rej</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 9 Results for the models in Table 8 on information measures from mutual-information and cross-entropy groups ((R) = ranking order for the model, where R = A, B, ..., in descending order from the top)

<table>
<thead>
<tr>
<th>Model/(R)</th>
<th>N1</th>
<th>N2</th>
<th>N3</th>
<th>N4</th>
<th>N5</th>
<th>N6</th>
<th>N7</th>
<th>N8</th>
<th>N9</th>
<th>N10</th>
</tr>
</thead>
<tbody>
<tr>
<td>M7</td>
<td>0.912</td>
<td>0.912</td>
<td>0.957</td>
<td>0.935</td>
<td>0.934</td>
<td>0.934</td>
<td>0.876</td>
<td>0.912</td>
<td>0.957</td>
<td>0.998</td>
</tr>
<tr>
<td>(F)</td>
<td>(F)</td>
<td>(F)</td>
<td>(C)</td>
<td>(G)</td>
<td>(G)</td>
<td>(F)</td>
<td>(H)</td>
<td>(E)</td>
<td>(D)</td>
<td>(C)</td>
</tr>
<tr>
<td>M8</td>
<td>0.939</td>
<td>0.939</td>
<td>0.958</td>
<td>0.949</td>
<td>0.949</td>
<td>0.949</td>
<td>0.902</td>
<td>0.939</td>
<td>0.958</td>
<td>0.988</td>
</tr>
<tr>
<td>(F)</td>
<td>(E)</td>
<td>(E)</td>
<td>(D)</td>
<td>(D)</td>
<td>(D)</td>
<td>(D)</td>
<td>(D)</td>
<td>(D)</td>
<td>(C)</td>
<td>(C)</td>
</tr>
<tr>
<td>M9</td>
<td>1.000</td>
<td>0.951</td>
<td>0.961</td>
<td>0.980</td>
<td>0.980</td>
<td>0.980</td>
<td>0.961</td>
<td>0.961</td>
<td>1.000</td>
<td>0.982</td>
</tr>
<tr>
<td>(C)</td>
<td>(A)</td>
<td>(D)</td>
<td>(A)</td>
<td>(A)</td>
<td>(A)</td>
<td>(A)</td>
<td>(A)</td>
<td>(A)</td>
<td>(G)</td>
<td>(G)</td>
</tr>
<tr>
<td>M10</td>
<td>0.912</td>
<td>0.912</td>
<td>0.938</td>
<td>0.925</td>
<td>0.925</td>
<td>0.925</td>
<td>0.860</td>
<td>0.912</td>
<td>0.938</td>
<td>0.999</td>
</tr>
<tr>
<td>(E)</td>
<td>(F)</td>
<td>(F)</td>
<td>(F)</td>
<td>(I)</td>
<td>(I)</td>
<td>(I)</td>
<td>(I)</td>
<td>(I)</td>
<td>(H)</td>
<td>(G)</td>
</tr>
<tr>
<td>M11</td>
<td>0.956</td>
<td>0.956</td>
<td>0.941</td>
<td>0.948</td>
<td>0.948</td>
<td>0.948</td>
<td>0.902</td>
<td>0.941</td>
<td>0.956</td>
<td>0.998</td>
</tr>
<tr>
<td>(E)</td>
<td>(D)</td>
<td>(C)</td>
<td>(E)</td>
<td>(E)</td>
<td>(E)</td>
<td>(E)</td>
<td>(D)</td>
<td>(C)</td>
<td>(E)</td>
<td>(B)</td>
</tr>
<tr>
<td>M12</td>
<td>1.000</td>
<td>0.969</td>
<td>0.943</td>
<td>0.972</td>
<td>0.971</td>
<td>0.971</td>
<td>0.943</td>
<td>0.943</td>
<td>1.000</td>
<td>0.983</td>
</tr>
<tr>
<td>(B)</td>
<td>(A)</td>
<td>(B)</td>
<td>(D)</td>
<td>(B)</td>
<td>(B)</td>
<td>(B)</td>
<td>(B)</td>
<td>(A)</td>
<td>(F)</td>
<td>(G)</td>
</tr>
<tr>
<td>M13</td>
<td>0.939</td>
<td>0.939</td>
<td>0.915</td>
<td>0.927</td>
<td>0.927</td>
<td>0.927</td>
<td>0.863</td>
<td>0.915</td>
<td>0.939</td>
<td>0.999</td>
</tr>
<tr>
<td>(D)</td>
<td>(E)</td>
<td>(E)</td>
<td>(I)</td>
<td>(H)</td>
<td>(H)</td>
<td>(H)</td>
<td>(H)</td>
<td>(G)</td>
<td>(F)</td>
<td>(A)</td>
</tr>
<tr>
<td>M14</td>
<td>0.956</td>
<td>0.956</td>
<td>0.916</td>
<td>0.936</td>
<td>0.935</td>
<td>0.935</td>
<td>0.879</td>
<td>0.916</td>
<td>0.956</td>
<td>0.998</td>
</tr>
<tr>
<td>(A)</td>
<td>(A)</td>
<td>(A)</td>
<td>(G)</td>
<td>(C)</td>
<td>(C)</td>
<td>(C)</td>
<td>(C)</td>
<td>(E)</td>
<td>(A)</td>
<td>(E)</td>
</tr>
<tr>
<td>M15</td>
<td>1.000</td>
<td>0.996</td>
<td>0.919</td>
<td>0.960</td>
<td>0.958</td>
<td>0.959</td>
<td>0.919</td>
<td>0.919</td>
<td>1.000</td>
<td>0.984</td>
</tr>
<tr>
<td>(A)</td>
<td>(A)</td>
<td>(A)</td>
<td>(G)</td>
<td>(C)</td>
<td>(C)</td>
<td>(C)</td>
<td>(C)</td>
<td>(E)</td>
<td>(A)</td>
<td>(E)</td>
</tr>
</tbody>
</table>
from both theoretical and numerical ones, were made on this promising measure. The theoretical properties of this measure were derived in Appendix. While Theorem A1 confirms that \( N_{I2} \) satisfies Feature 3 around the exact problems. Although some \( N_1 \)s could be removed directly based on their poor performance with respect to the meta-measures (such as \( N_{I1} \) and \( N_{I0} \) on Feature 2), they were retained to demonstrate pros and cons in the applications. At this stage, we extend the concepts of error types and reject types to multiple classes. Nine examples are specifically designed in Table 8. The ranking order for each model is shown in Table 8, which is derived from the principles of Feature 3. From Tables 9 and 10, it is interesting to see that \( N_{I0} \) is still the most appropriate measure for classification evaluations. Using this measure, we can select the “best” and “worst” classifications consistent with our intuition. All other measures cannot distinguish error types and reject types properly. The examples in Tables 4, 7, and 8 only present limited scenarios for variations in confusion matrices. Using the open-source toolbox from [53], one is able to test more scenarios for numerical investigations.

Table 11 demonstrates comparisons of the objective evaluation measures from both performance and information categories in relation to the meta-measures. None of them is able to test more scenarios for numerical investigations. The meta-measures provide users with a simple guideline of selecting “proper” measures for their specific concerns of applications. For example, the performance measures satisfy Feature 1, but fail to directly distinguish error types and reject types in an objective evaluation. When Feature 2 or 3 is a main concern, the information measures (such as KL divergence), for the given data.

Further investigations were carried out on three-class problems. Although some \( N_1 \)s could be removed directly based on their poor performance with respect to the meta-measures (such as \( N_{I1} \) and \( N_{I0} \) on Feature 2), they were retained to demonstrate pros and cons in the applications. At this stage, we extend the concepts of error types and reject types to multiple classes. Nine examples are specifically designed in Table 8. The ranking order for each model is shown in Table 8, which is derived from the principles of Feature 3. From Tables 9 and 10, it is interesting to see that \( N_{I0} \) is still the most appropriate measure for classification evaluations. Using this measure, we can select the “best” and “worst” classifications consistent with our intuition. All other measures cannot distinguish error types and reject types properly. The examples in Tables 4, 7, and 8 only present limited scenarios for variations in confusion matrices. Using the open-source toolbox from [53], one is able to test more scenarios for numerical investigations.

Table 11 demonstrates comparisons of the objective evaluation measures from both performance and information categories in relation to the meta-measures. None of them is able to test more scenarios for numerical investigations. The meta-measures provide users with a simple guideline of selecting “proper” measures for their specific concerns of applications. For example, the performance measures satisfy Feature 1, but fail to directly distinguish error types and reject types in an objective evaluation. When Feature 2 or 3 is a main concern, the information measures (such as KL divergence), for the given data.

Further investigations were carried out on three-class problems. Although some \( N_1 \)s could be removed directly based on their poor performance with respect to the meta-measures (such as \( N_{I1} \) and \( N_{I0} \) on Feature 2), they were retained to demonstrate pros and cons in the applications. At this stage, we extend the concepts of error types and reject types to multiple classes. Nine examples are specifically designed in Table 8. The ranking order for each model is shown in Table 8, which is derived from the principles of Feature 3. From Tables 9 and 10, it is interesting to see that \( N_{I0} \) is still the most appropriate measure for classification evaluations. Using this measure, we can select the “best” and “worst” classifications consistent with our intuition. All other measures cannot distinguish error types and reject types properly. The examples in Tables 4, 7, and 8 only present limited scenarios for variations in confusion matrices. Using the open-source toolbox from [53], one is able to test more scenarios for numerical investigations.

Table 11 demonstrates comparisons of the objective evaluation measures from both performance and information categories in relation to the meta-measures. None of them is able to test more scenarios for numerical investigations. The meta-measures provide users with a simple guideline of selecting “proper” measures for their specific concerns of applications. For example, the performance measures satisfy Feature 1, but fail to directly distinguish error types and reject types in an objective evaluation. When Feature 2 or 3 is a main concern, the information measures (such as KL divergence), for the given data.

Further investigations were carried out on three-class problems. Although some \( N_1 \)s could be removed directly based on their poor performance with respect to the meta-measures (such as \( N_{I1} \) and \( N_{I0} \) on Feature 2), they were retained to demonstrate pros and cons in the applications. At this stage, we extend the concepts of error types and reject types to multiple classes. Nine examples are specifically designed in Table 8. The ranking order for each model is shown in Table 8, which is derived from the principles of Feature 3. From Tables 9 and 10, it is interesting to see that \( N_{I0} \) is still the most appropriate measure for classification evaluations. Using this measure, we can select the “best” and “worst” classifications consistent with our intuition. All other measures cannot distinguish error types and reject types properly. The examples in Tables 4, 7, and 8 only present limited scenarios for variations in confusion matrices. Using the open-source toolbox from [53], one is able to test more scenarios for numerical investigations.
7 Summary

In this work, we investigated objective evaluations of classifications by introducing normalized information measures. We reviewed the related works and discussed objectivity and its formal definition in evaluations. Objective evaluations may be required under different application background. In classifications, for example, exact knowledge of misclassification costs is sometimes unknown in evaluations. Moreover, cases of ignorance regarding reject costs appear more often in scenarios of abstaining classifications. In these cases, although subjective evaluations can be applied, the user-given data of the unknown abstention costs will lead to a much higher degree of uncertainty or inconsistency. We believe that an objective evaluation can be an initial basis, or a complementary approach, to subjective evaluations. In some situations, an objective evaluation is considered useful despite the subjective evaluations being reasonable for the applications. The results from both objective and subjective evaluations give users an overall quality of classification results.

Considering that abstaining classifications are becoming more popular, we focused on the distinctions of error types and reject types in objective evaluations of classifications. First, we proposed three meta-measures for assessing classifications, which seem more relevant and proper than the objective evaluations. We reviewed the related works and discussed objective evaluations may be required under different application background. In classifications, for example, exact knowledge of misclassification costs is sometimes unknown in evaluations. Moreover, cases of ignorance regarding reject costs appear more often in scenarios of abstaining classifications. In these cases, although subjective evaluations can be applied, the user-given data of the unknown abstention costs will lead to a much higher degree of uncertainty or inconsistency. We believe that an objective evaluation can be an initial basis, or a complementary approach, to subjective evaluations. In some situations, an objective evaluation is considered useful despite the subjective evaluations being reasonable for the applications. The results from both objective and subjective evaluations give users an overall quality of classification results.

Considering that abstaining classifications are becoming more popular, we focused on the distinctions of error types and reject types in objective evaluations of classifications. First, we proposed three meta-measures for assessing classifications, which seem more relevant and proper than the objective evaluations. We reviewed the related works and discussed objective evaluations may be required under different application background. In classifications, for example, exact knowledge of misclassification costs is sometimes unknown in evaluations. Moreover, cases of ignorance regarding reject costs appear more often in scenarios of abstaining classifications. In these cases, although subjective evaluations can be applied, the user-given data of the unknown abstention costs will lead to a much higher degree of uncertainty or inconsistency. We believe that an objective evaluation can be an initial basis, or a complementary approach, to subjective evaluations. In some situations, an objective evaluation is considered useful despite the subjective evaluations being reasonable for the applications. The results from both objective and subjective evaluations give users an overall quality of classification results.

### Appendix. Theorems and sensitivity functions of NI2 for binary classifications

#### Theorem A1. For a binary classification defined by

\[
C = \begin{bmatrix}
TN & FP \\
FN & TP
\end{bmatrix}
\]

and

\[
C_1 = TN + FP + RN, C_2 = FN + TP + RP, C_1 + C_2 = n
\]

\(NI_2\) satisfies Feature 3 on the property regarding error types and reject types around the exact classifications. Specifically for the four confusion matrices below:

\[
M_1 = \begin{bmatrix}
C_1 & 0 & 0 \\
C_2 - d & 0 & 0
\end{bmatrix}, \quad M_2 = \begin{bmatrix}
C_1 - d & d & 0 \\
0 & C_2 & 0
\end{bmatrix}
\]

\[
M_3 = \begin{bmatrix}
C_1 & 0 & 0 \\
0 & C_2 - d & 0
\end{bmatrix}, \quad M_4 = \begin{bmatrix}
C_1 - d & 0 & d \\
0 & C_2 & 0
\end{bmatrix}
\]

The following relations hold:

\[
NI_2(M_1) < NI_2(M_2), \quad NI_2(M_3) < NI_2(M_4)
\]

\[
C_1 > C_2 \Rightarrow NI_2(M_1) > NI_2(M_2)
\]

#### Proof. For a binary classification, \(NI_2\) is defined by the modified mutual information \(I_M\):

\[
I_M(T,Y) = \frac{1}{n} \log_2 \frac{C_1(TN + FN)}{C_1 + d} + \frac{FP}{n} \log_2 \frac{C_1 + d}{C_1(TP + FP)} + \frac{FN}{n} \log_2 \frac{C_2 + d}{C_2(TN + FN)} + \frac{FP}{n} \log_2 \frac{C_2 + d}{C_2(TP + FP)}
\]

Let \(M_0\) be a confusion matrix corresponding to the exact classifications:

\[
M_0 = \begin{bmatrix}
C_1 & 0 & 0 \\
0 & C_2 & 0
\end{bmatrix}
\]

and be a baseline, one can obtain the analytical results below for the mutual information differences between the models:

\[
\Delta I_{10} = I_M(M_1) - I_M(M_0) = \frac{1}{n} \left( C_1 \log_2 \frac{C_1}{C_1 + d} + d \log_2 \frac{d}{C_1 + d} \right)
\]

\[
\Delta I_{20} = I_M(M_2) - I_M(M_0) = \frac{1}{n} \left( C_2 \log_2 \frac{C_2}{C_2 + d} + d \log_2 \frac{d}{C_2 + d} \right)
\]

\[
\Delta I_{30} = I_M(M_3) - I_M(M_0) = \frac{d}{n} \log_2 \frac{C_2}{C_2 + n}
\]

\[
\Delta I_{40} = I_M(M_4) - I_M(M_0) = \frac{d}{n} \log_2 \frac{C_1}{C_1 + n}
\]

For the given assumption \(C_1 > C_2 > d > 0\), all \(\Delta I_{\alpha}\) above are negative and we denote their absolute values to be the "\(\Delta I_{\alpha}\) costs in classifications." One can directly prove that \(\Delta I_{\alpha}\) costs in classifications. One can directly prove that \(\Delta I_{\alpha}\) costs in classifications. One can directly prove that \(\Delta I_{\alpha}\) costs in classifications.

\[
g_1(x) = \left( \frac{x}{x + d} \right)^2, \quad g_2(x) = \left( \frac{d}{x + d} \right)^d, \quad x > 0, d > 0
\]

Then, from the proof above, one can derive the following relations:

\[
C_1 > C_2 \Rightarrow \left( \frac{d}{C_1 + d} \right)^d < \left( \frac{d}{C_2 + d} \right)^d < 1
\]

#### 7.1 Acknowledgments

The editorial assistance for improving the manuscript from Mr. Christian Ocier is gratefully acknowledged.
The relations in (A3a) are achieved for $NI_2$ because its normalization term, $H(T_1)$, is a constant for the given $C_1$ and $C_2$. One therefore confirms the satisfaction of Feature 3 on the property of the within error types and reject types around the exact classifications, respectively.

Then, there is a proof of the relation (A8b), which suggests that a misclassification suffers a higher cost than a rejection for the same class. Feature 3 considers this relation as a basic property in classifications for the between error and reject types. The procedures for the proof are:

$$C_1 > C_2 \rightarrow C_1C_2 + C_1d > (C_1 + C_2)d = nd \rightarrow$$

$$1 > C_1 \frac{d}{C_1 + d} \rightarrow \left| \log_2 \left( \frac{C_1}{n} \right) \right| < \left| \log_2 \left( \frac{d}{C_2 + d} \right) \right| \rightarrow$$

$$\frac{1}{n} \left| d \log_2 \left( \frac{C_1}{n} \right) \right| < \frac{1}{n} \left| d \log_2 \left( \frac{d}{C_2 + d} \right) \right| \rightarrow$$

$$\frac{1}{n} \left| C_1 \log_2 \left( \frac{C_2}{C_1 + d} + d \log_2 \frac{d}{C_2 + d} \right) \right| \rightarrow$$

$$|\Delta I_{20}| < |\Delta I_{20}|$$

(A8a)

$$C_1 + d < n \rightarrow C_1(C_1 + d) + nd < C_1n + nd \rightarrow$$

$$\frac{C_1(C_1 + d) + nd}{n(C_1 + d)} < 1 \rightarrow \frac{C_1}{n} + \frac{d}{C_2 + d} < 1 \rightarrow$$

$$\frac{d}{C_2 + d} < \frac{C_2}{n} < 1 \rightarrow \left| \log_2 \left( \frac{C_2}{n} \right) \right| < \left| \log_2 \left( \frac{d}{C_2 + d} \right) \right| \rightarrow$$

$$\frac{1}{n} \left| d \log_2 \left( \frac{C_2}{n} \right) \right| < \frac{1}{n} \left| C_1 \log_2 \left( \frac{C_1}{C_1 + d} + d \log_2 \frac{d}{C_2 + d} \right) \right| \rightarrow$$

$$|\Delta I_{20}| < |\Delta I_{20}|$$

(A8b)

Theorem A2. For the given conditions (A1) and (A2) and $C_1 > C_2 > d > 0$, $NI_2$ will satisfy the following relations:

$$NI_2(M_4) > NI_2(M_3) > NI_2(M_2) > NI_2(M_1),$$

$$0.5 < p_1 < p_2 \leq 1$$

(A9a)

$$NI_2(M_4) > NI_2(M_3) > NI_2(M_2) > NI_2(M_1),$$

$$0.5 < p_1 < p_2 \leq 1$$

(A9b)

if a critical boundary $p_c$ exists and we set $p_1 = C_1/n$.

Proof. The critical boundary, $p_c$, is determined by the cross-over point between the functions of (A6b) and (A6c), or, from solving the equation:

$$f = \Delta I_{20} - \Delta I_{20} =$$

$$\frac{1}{n} \left( C_1 \log_2 \left( \frac{C_2}{C_2 + d} + d \log_2 \frac{d}{C_2(C_2 + d)} \right) \right) = 0$$

(A10)

There exists no closed-form solution to $p_c$. Monotonically increasing and decreasing functions of (A6b) and (A6c) enable only a single cross-over point in the region of $p_1 > 0.5$. Based on the monotonicity of the functions and relations in (A3), one is able to confirm the conditions in (A9a) and (A9b), respectively.

Fig. A1 depicts the case when $d = 1$ and $n = 100$.}

**Remark A1.** The value of $p_c$ is inversely proportional to the independent variable of $d/n$. A numerical solution to $p_c$ should be engaged. The physical interpretation of $p_c$ is a critical point at which a misclassification from a large class has the same cost as with a rejection from a small class. The plots of $\Delta I_{20}$ and $\Delta I_{20}$ reveal a novel finding about “which costs more, a misclassification from a large class or a rejection from a small class?”

The finding confirms that information theory is principally general to deal with errors, rejects, and their relations.

The sensitivity functions are given as the conventional forms for delivering approximation analysis of $I_{2n}$:

$$\frac{\partial I_{1n}}{\partial T_N} = - \frac{1}{n} \left[ \log_2 \left( \frac{C_1}{n} \right) + \left( \log_2 \left( \frac{TN}{TN + FN} \right) \right) \right]$$

(A11a)

$$\frac{\partial I_{1n}}{\partial T_P} = - \frac{1}{n} \left[ \log_2 \left( \frac{C_2}{n} \right) + \left( \log_2 \left( \frac{TP}{TP + FP} \right) \right) \right]$$

(A11b)

$$\frac{\partial I_{1n}}{\partial F_N} = - \frac{1}{n} \left[ \log_2 \left( \frac{C_2}{n} \right) + \left( \log_2 \left( \frac{FN}{FN + TN} \right) \right) \right]$$

(A11c)

$$\frac{\partial I_{1n}}{\partial F_P} = - \frac{1}{n} \left[ \log_2 \left( \frac{C_1}{n} \right) + \left( \log_2 \left( \frac{FP}{FP + TP} \right) \right) \right]$$

(A11d)

$$\frac{\partial I_{1n}}{\partial R_N} = - \frac{\partial I}{\partial TN} - \frac{\partial I}{\partial FP}$$

(A11e)

$$\frac{\partial I_{1n}}{\partial R_P} = - \frac{\partial I}{\partial FN} - \frac{\partial I}{\partial TP}$$

(A11f)

Only four independent variables describe the sensitivity functions due to the two constraints in (A11b). Hence, a chain rule is applied for deriving the functions of (A11e) and (A11f).

**Remark A2.** Using (A11), we failed to reach the reasonable conclusions as those in Theorems A1 for the reason that the first-order differentials may be not sufficient for the analysis around the exact classifications. For example, we got the results:

$$I(M_1) - I(M_0) \approx$$

$$\frac{T_P}{S_1} \left[ I(M_0) + (FN_1 - FN_0) \frac{\partial I(M_0)}{\partial FN} \right] =$$

$$- \frac{d}{n} \log_2 \left( \frac{p_2}{c_2} \right) \left( 0 \right) = 0$$

(A12a)

$$I(M_2) - I(M_0) \approx$$

$$\frac{T_P}{S_1} \left[ I(M_0) + (FN_1 - FN_0) \frac{\partial I(M_0)}{\partial FN} \right] =$$

$$- \frac{d}{n} \log_2 \left( \frac{p_2}{c_1} \right) \left( 0 \right) = 0$$

(A12b)

This observation suggests that one needs to be cautious when using sensitivity function for approximation analysis on $I_{1n}$ (or $NI_2$).
References


41 Berger J O. The case for objective Bayesian analysis. Bayesian Analysis, 2006, 1(3): 385–402


HU Bao-Gang Professor at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, P. R. China. He received his Ph.D. degree in mechanical engineering from McMaster University, Canada in 1993. From 1994 to 1997, he was a research engineer and senior research engineer at C-CORE, Memorial University of Newfoundland, Canada. His research interest covers intelligent systems, pattern recognition, and plant growth modeling. Corresponding author of this paper. E-mail: hubg@nlpr.ia.ac.cn

HE Ran Assistant professor in the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, P. R. China. He received his Ph.D. degree in pattern recognition and intelligent systems from Institute of Automation, Chinese Academy of Sciences, P. R. China in 2009. His research interest covers information theoretic learning, pattern recognition, and computer vision. E-mail: rhe@nlpr.ia.ac.cn

YUAN Xiao-Tong Postdoctoral fellow in the Department of Statistics and Bio-statistics, Rutgers University, USA. He received his Ph. D. degree in pattern recognition and intelligent system from Institute of Automation, Chinese Academy of Sciences, P. R. China in 2009. His research interest covers machine learning, data mining, and computer vision. E-mail: xtyuan1980@gmail.com