Acta Automatica Sinica
Vol. 34, No. 11 November, 2008

Evaluation Criteria Based on Mutual Information for Classifications Including Rejected Class

HU Bao-Gang[1, 2] WANG Yong[1, 2]

Abstract Different from the conventional evaluation criteria using performance measures, information theory based criteria present a unique beneficial feature in applications of machine learning. However, we are still far from possessing an in-depth understanding of the "entropy" type criteria, say, in relation to the conventional performance-based criteria. This paper studies generic classification problems, which include a rejected, or unknown, class. We present the basic formulas and schematic diagram of classification learning based on information theory. A closed-form equation is derived between the normalized mutual information and the augmented confusion matrix for the generic classification problems. Three theorems and one set of sensitivity equations are given for studying the relations between mutual information and conventional performance indices. We also present numerical examples and several discussions related to advantages and limitations of mutual information criteria in comparison with the conventional criteria.

Key words Entropy, mutual information, evaluation criteria, classification, confusion matrix, machine learning

It is understandable that evaluation criteria (sometimes equivalent to learning targets) comprise the first task in studies on machine learning. This task will be simplified if some criteria are specified with the application requirements. However, from the theoretical point of view, selections of evaluation criteria are nevertheless an open problem in machine learning. For better understanding of the problem, we roughly categorize evaluation criteria into several groups. Within a type of performance-based criteria, taking classification problems for examples, one can further divide them as partial performance criterion like "true positive accuracy", or overall performance one like ROC (Receiver operating characteristic) curves. This type of performance-based criteria can still be grouped as direct measure criteria or indirect measure ones. The direct measure criteria include classification error, ROC curve, or computational cost. The indirect measure criteria can be found as mutual information, class separation margins, or error bounds. Up to now, most selections of learning criteria are made based on users' experiences or preferences. Therefore, a systematic study seems to be necessary in order to explore the subject, including the following two basic issues:

1) One of the principal tasks in machine learning is to process data. Can we apply "entropy" or information-based criteria as a generic measure for dealing with uncertainty of data in machine learning?

2) What are the relations between information-based criteria (say, mutual information) and the conventional performance criteria (say, classification accuracy)? What are the advantages and limitations in using information-based criteria?

This paper will address the second issue. Considering that information-based criteria have been extensively applied in the studies of unsupervised learning[1−2], we will focus on supervised learning, particularly on generic classification problems which include a rejected class. The main objectives of this work are to derive new formulas of normalized mutual information from the augmented confusion matrix and to provide theoretical interpretations of mutual information criteria in the context of classification problems. At the same time, some numerical examples and computer program are given. Finally, we summarize the advantages and limitations of mutual information criteria in classification problems.

Received November 6, 2007; in revised form August 14, 2008
Supported by National Natural Science Foundation of China (60275025, 60121302)

1. National Laboratory of Pattern Recognition /Sino-French Laboratory in Computer Science, Automation and Applied Mathematics, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, P. R. China 2. Graduate University of Chinese Academy of Sciences, Beijing 100049, P. R. China

1 Related work

In this section, we introduce some existing works related to the mutual information criteria. At the same time, some basic equations are given so that readers can follow the new formulas derived in the next section. Shannon introduced "entropy" concept into information theory as[3]

$$H(X) = -\sum p(x)\log_2 p(x) \tag{1}$$

where X is a discrete random variable with probability density function (PDF) $p(x)$. As the entropy is considered as a measure of uncertainty of random variable, it is also viewed as a measure of impurity in data[4]. The mutual information is defined as the relative entropy, or mutual entropy[5], between the joint distribution $p(x, y)$ and the product distribution

$$I(X, Y) = \sum\sum p(x, y)\log_2 \frac{p(x, y)}{p(x)p(y)} \tag{2}$$

where X and Y are two random variables, respectively. The relations between the entropy and mutual information are:

$$I(X, Y) = I(Y, X) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{3}$$

$$I(X, X) = H(X) \tag{4}$$

where $H(Y|X)$ is the conditional entropy, and is defined as

$$H(Y|X) = \sum p(x)H(Y|X = x) \tag{5}$$

Equation (4) shows that any entropy can be considered as self mutual information. Fig. 1 shows the relationships of (3) in a context of classifications, where X is replaced by a target variable T. It was reported[1] that the proposal of applying mutual information criteria in machine learning is mostly attributed to the study of Linsker[6]. After his work, significant investigations have been reported on the subjects of feature selection/extraction[7−10], independent component analysis[11], image registration[12−13], etc. Haykin[1] systematically summarized the four cases of using mutual information as an objective function in neural network studies. Xu[2] and Principe[14] proposed a generic framework in machine learning for both supervised learning and unsupervised learning.

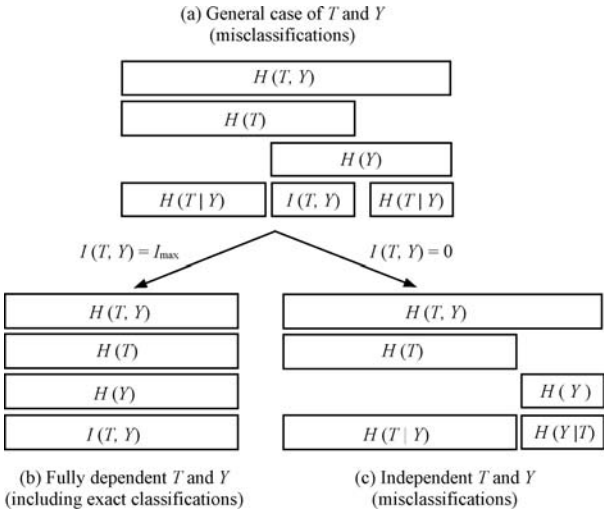

Fig. 1 Relationships between mutual information and entropy in a context of classification problems, where $H(T)$ is generally fixed as a baseline ((a) General case[5] (misclassifications); (b) Fully dependent T and Y (including exact classifications); (c) Independent T and Y (misclassifications))

In fact, there exist numerous definitions of information-based criteria in applications[14]. Among them, one important notation is the normalized mutual information[3, 15], defined as

$$NI(X,Y) = \frac{I(X,Y)}{H(X)}, \quad NI(Y,X) = \frac{I(Y,X)}{H(Y)} \tag{6}$$

This notation presents another type of "correlation" measure[3] and sometimes is called as "asymmetric dependency coefficient (ADC)"[15]. However, two definitions in (6) will produce unequal values due to their asymmetric property in the definitions. Therefore, [16−17] proposed normalized mutual information with symmetric property, such as

$$NI(X,Y) = 2\frac{I(X,Y)}{H(Y)+H(X)}, NI(X,Y) = \frac{I(X,Y)}{\sqrt{H(Y)H(X)}} \tag{7}$$

Quinlan[18−19] proposed a new term called "information gain" as a criterion for the study of decision tree. The basic formula of this criterion is given in the following definition by Mitchell[4]

$$IG(S,A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \tag{8}$$

where S and A represent the sets of samples and attributes, respectively, and $|\cdot|$ is a frequency calculation. The first part on the right side of (8) is entropy with respect to the sample labels, rather than to the sample data themselves. The second part is the conditional entropy with respect to the predication labels. This information gain criterion presents a principle of forming a decision tree with a simple structure. However, with the explanation by Quinlan[19], "information gain is also known as the mutual information between the test X and the class". Ding and Wu[20] proposed a term of "entropy decrement of recognition", which has the same meaning as mutual information. They also derived three theorems for studies of entropy-based pattern recognitions.

In the study on the relationship between mutual information and Bayesian errors, Fano's pioneer work[21] resulted in the following condition

$$\Pr(Y \neq T) \geq \frac{H(T) - I(T,Y) - 1}{\log_2(m)} = \frac{H(T|Y) - 1}{\log_2(m)} \tag{9}$$

where $\Pr(Y \neq T)$ is a Bayesian error and m is a total number of classes, but notation T is used for classification problems. This inequality presents a lower bound for the Bayesian error. And, the upper bound was derived by Hellman and Raviv[22]

$$\Pr(Y \neq T) \leq \frac{H(T) - I(T,Y)}{2} = \frac{H(T|Y)}{2} \tag{10}$$

Eriksson[23] pointed out that (10) should be effective only when $\Pr(Y \neq T) \leq 0.5$. In general, it seems more important that one can have an exact relation between mutual information and classification accuracy. In a recent work, we derived the nonlinear relations between normalized mutual information and the conventional criteria for binary classification problems[24]. The present work is its extension of the analysis to classifications on multiple classes by including a rejected class as a generic approach.

2 Definitions and formulas for classification problems

Definition 1. A generic classification problem is defined as a classification, which may assign samples into a rejected, or unknown, class. Therefore, the three data sets that are used for training a classifier will be defined as: input data set $\{\boldsymbol{x}_k\}_{k=1}^n \in X \subset \mathbf{R}^d$; output data set $\{y_k\}_{k=1}^n \in Y = \{1, 2, \cdots, m+1\}$; and target data set $\{t_k\}_{k=1}^n \in T = \{1, 2, \cdots, m\}$, respectively; where n is a total sample number, m is a total class number, and d is the dimensions of feature space. When $y_k = m+1$, it represents a rejected class.

Remark 1. Different from X in feature space that could be any type of data, both T and Y represent label information as integer sets for hard classifiers. In real applications, a rejection strategy is often used for improving the accuracy of classifiers. For this reason, we define Y to have one more label than T for a rejected class.

Definition 2. For a generic classification problem, an augmented confusion matrix is defined by adding one column for a rejected class onto a conventional confusion matrix in a form as

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

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 exact classes, and the column data corresponds to the prediction classes. The last column represents a rejected class. The constraints of an augmented confusion matrix are

$$C_i = \sum_{j=1}^{m+1} c_{ij}, \ C_i > 0, \ c_{ij} \geq 0, \ i = 1, 2, \cdots, m \tag{12}$$

where C_i is the total number for the i-th class. The data for C_i is known in classification problems.

Definition 3. The empirical PDF of joint distribution $P_e(T,Y)$ in classification problems is defined from a frequency means on C as

$$P_e(T,Y) = \left(\frac{c_{ij}}{n}\right)_{m\times(m+1)}$$
$$i = 1,2,\cdots,m,\ j = 1,2,\cdots,m+1 \tag{13}$$

where $n = \sum C_i$ is a known constant in classification problems. Then, the empirical PDFs for the marginal distributions are

$$P_e(T) = \left(\frac{C_i}{n}\right)_{m\times 1},\quad i = 1,2,\cdots,m \tag{14}$$

$$P_e(Y) = \left(\frac{1}{n}\sum_{i=1}^{m} c_{ij}\right)_{1\times(m+1)},\quad j = 1,2,\cdots,m+1 \tag{15}$$

Remark 2. When the exact PDFs for Y and T are unknown, one can use the empirical ones for an approximation study. Table 1 lists all terms in (13) ∼ (15). This table will be helpful for understanding and calculating the empirical information-based criteria in classification problems.

Modified on the studies by Xu[2] and Principe[14], we present a schematic diagram of classification learning based on information theory (see Fig. 2). The objective function is set as

$$\max\ NI(T,Y) = NI(T, f(X,\boldsymbol{\theta})) = \frac{I(T,Y)}{H(T)} \tag{16}$$

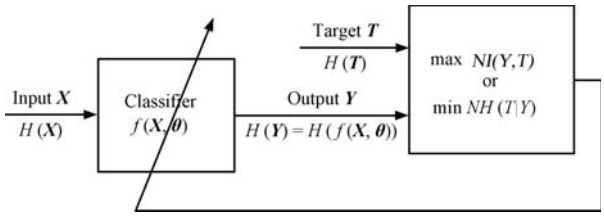

Fig. 2 Schematic diagram of classifications based on information theory

The basic interpretations for this objective function are to design the nonlinear function f, and then, to tune the parameter vector $\boldsymbol{\theta}$ for maximizing the correlations between Y and T. When $NI = 1$, it represents a full correlation between Y and T (see Fig. 1 (b)). When $NI = 0$, it indicates complete independence between Y and T (see Fig. 1 (c)). From Fig. 1, one can see that an exact classification is a process to force $I(T,Y)$ or $H(Y)$ to reach the baseline of $H(T)$, which is generally unchanged. Theoretically, one can find out that minimization of normalized conditional entropy

$$\min\ NH(T|Y) = \frac{H(T|Y)}{H(T)} \tag{17}$$

will be fully equivalent to (16). The interpretation for this objective function is meant to minimize the uncertainty of T when Y is given. Note that we apply the asymmetric definitions on $NI(T,Y)$ and $NH(T|Y)$ for the consideration of simplicity. Within (16) and (17), $H(T)$ is usually known and fixed for representing the uncertainty of the target data. Contrarily, $H(Y)$ needs to be updated in each iteration of learning, which will add extra computational cost.

In general, when the exact PDFs are unknown, the empirical definitions of entropy and mutual information have to be used. The empirical entropy of target data is then given by

$$H_e(T) = -\sum_{i=1}^{m}\frac{C_i}{n}\log_2\left(\frac{C_i}{n}\right) \tag{18}$$

We use the following form in calculating the empirical mutual information

$$I_e(T,Y) = \sum_{T}\sum_{Y} P_e(T,Y)\log_2\frac{P_e(T,Y)}{P_e(T)P_e(Y)} \tag{19}$$

From Table 1 and substitutions of (13) ∼ (15) and (18) ∼ (19) into NI expression in (16), one is convenient to obtain the empirically normalized mutual information as

$$NI_e(T,Y) = \frac{I_e(T,Y)}{H_e(T)} = -\frac{\displaystyle\sum_{i=1}^{m}\sum_{j=1}^{m+1} c_{ij}\log_2\left(\frac{c_{ij}}{C_i\displaystyle\sum_{i=1}^{m}\left(\frac{c_{ij}}{n}\right)}\right)}{\displaystyle\sum_{i=1}^{m} C_i\log_2\left(\frac{C_i}{n}\right)} \tag{20}$$

Remark 3. Equation (20) presents a generic formula for calculating the empirically normalized mutual information in classification problems, which includes cases for assigning samples into a rejected class. When using this equation, a specific operation should be made on removing the singularity points in numerical studies, which is for imposing the similar condition $H(0) = 0$ [3] on (20).

3 Relations between mutual information and conventional performance measures

In this section, we will address the second open issue discussed in the introduction. The knowledge about the relations between mutual information and the conventional performance measures is fundamental from both theoretical

Table 1 Empirical PDFs of the joint distribution $P_e(T,Y)$, and the marginal distributions $P_e(Y)$, $P_e(T)$ in classification problems

$P_e(T,Y)$		Y					$P_e(T)$
		1	2	⋯	m	$m+1$	
T	1	c_{11}/n	c_{12}/n	⋯	c_{1m}/n	$c_{1(m+1)}/n$	C_1/n
	2	c_{21}/n	c_{22}/n	⋯	c_{2m}/n	$c_{2(m+1)}/n$	C_2/n
	⋮	⋮	⋮	⋱	⋮	⋮	⋮
	m	c_{m1}/n	c_{m2}/n	⋯	c_{mm}/n	$c_{m(m+1)}/n$	C_m/n
$P_e(Y)$		$\sum_{i=1}^{m} c_{i1}/n$	$\sum_{i=1}^{m} c_{i2}/n$	⋯	$\sum_{i=1}^{m} c_{im}/n$	$\sum_{i=1}^{m} c_{i(m+1)}/n$	1

and application viewpoints. For the generic classification problems concerned in this work, we define accuracy, error rate, and rejection rate as

$$A = \frac{1}{n}\sum_{i=1}^{m} c_{ii}, E = \frac{1}{n}\sum_{i=1}^{m}\sum_{j\neq i}^{m} c_{ij}, Rej = \frac{1}{n}\sum_{i=1}^{m} c_{i(m+1)} \tag{21}$$

Note that in this generic case, the relation among them is

$$A + E + Rej = 1 \tag{22}$$

The other specific performance measures are appeared in binary classification problems (see Table 2). To follow the conventional terms in this case, we have a redefinition as

$$C = \begin{bmatrix} TP & FN & U_1 \\ FP & TN & U_2 \end{bmatrix} \tag{23}$$

where TP is for "true positive", FP for "false positive", FN for "false negative", TN for "true negative", and U_1 and U_2 for the rejected class from C_1 and C_2 samples, respectively.

Definition 4. In binary classifications, the performance measure described by single (or dual, or four) independent variable(s) is called as single- (or dual-, or four-) variable measure.

Remark 4. When binary classifications without cases for a rejected class are considered, one will have a confusion matrix formed by four variables. Because C_1 and C_2 are known as constraints, this matrix will have only two independent variables. In this situation, any performance measure can be described by at most two independent variables. When a rejected class is considered, a performance measure can be described by at most four independent variables.

Definition 5. In classifications, the overall performance measure is defined as a measure, which is able to present complete information about the classifier′s performance. For binary classifications without a rejected class, this measure has to be a dual-variable one and be evaluated from all operation points on the same classifier. If a rejected class is considered, the measure needs to be a four-variable one. Any partial performance measure provides only partial information of a classifier′s performance, which may not have sufficient variables or may not evaluate the classifier on all operation points.

Remark 5. Some measures can change their type depending on their use. Take accuracy measure for example. When this measure is applied to evaluate classifiers on a specific operation point, it is considered as a partial performance measure. The performance ranking, say, for two classifiers, can be changed on other operation points. If this measure is obtained in a principle of averaging accuracies on all operation points, it becomes an overall performance measure.

Remark 6. For binary classifications without a rejected class, performance evaluations on "True positive rate vs. False alarm rate"[25] or "Precision vs. Recall" curves will present an overall measure. However, when considering a rejected class, "Error rate vs. Rejection rate" curves[5] will be useful as an overall performance measure.

Remark 7. Classifications intrinsically present a conflicting property among some performance measures like the curves mentioned above. Therefore, a strategy of using ROC curves or AUC (Area under curve) index[26] will present an overall performance measure. It apparently remains an open issue to find a sensible and statistical measure for balancing the conflicting performance measures in classification problems.

Theorem 1. If a classifier assigns no sample into the rejected class, when $NI(T,Y) = 1$, or $H(T|Y) = 0$, the classifier corresponds to the cases of either an exact classification ($A = 1$) or a specific misclassification, which can be fully corrected by simple exchanges among labels. For a binary classifier, this misclassification is completely wrong ($A = 0$).

Proof. For a classifier without a rejected class, when $NI(T,Y) = 1$ or $H(T|Y) = 0$, one can obtain the following conditions from (20)

$$c_{ij} = C_i, c_{kj} = 0, i = 1, 2, \cdots, m, j = 1, 2, \cdots, m+1, k \neq i \tag{24}$$

These conditions indicate that within each of the first m columns, only one element, c_{ij}, equals to the class number C_i, and all other elements are zeros. When $j = i$ for all columns, it represents an exact classification. When $j \neq i$, it indicates there exists a zero on a diagonal element, which implies a misclassification. In this case, for a binary classification, the constraints (12) lead to $c_{12} = C_1$ and $c_{21} = C_2$, which exhibits a completely wrong result ($A = 0$) in the classification. From (24), one can see that the exact classification can be obtained by simple exchanges among labels for this type of misclassifications. □

Theorem 2. When $I(T,Y) = 0$, or $NH(T|Y) = 1$, the classifier exhibits a misclassification. One specific case is that all samples are considered to be one of m classes or a rejected class.

Table 2 Conventional performance measures and their formulas in binary classifications (C_1 is the number of exact positive, C_2 is the number of exact negative, n is the number of total samples, TP is the number of true positive, FP is the number of false positive, TN is the number of true negative, and FN is the number of false negative.)

Type	Independent variable (s)	Term	Formula
Partial performance	Single	True positive rate (TPR) (Recall, Hit rate)	TP/C_1
		True negative rate (TNR)	TN/C_2
		False alarm rate (FAR)	FP/C_2
	Dual	Precision (P)	$TP/(TP+FP)$
		Accuracy (A)	$(TP+TN)/n$
		Error rate (E)	$(FP+FN)/n$
Overall performance	Dual	Hit rate-false alarm rate (ROC curve)	Area under curve (AUC)
		Recall-precision (R-P curve)	Area under curve (AUC)
	Four	Error rate-rejection rate	Area under curve (AUC)

Proof. Apply a counter proof approach. Suppose that no error exists for the classifier, so that only the diagonal elements in C are non-zeros. Applying constraints (12) will result in $c_{ii} = C_i$. Hence, it will lead to $I(T,Y) = 1$ or $NH(T|Y) = 0$, which is against the original assumption in the theorem. Therefore, at least one non-diagonal element in C must be non-zero, which forms the following condition

$$\exists c_{ij} > 0, i \neq j, i = 1, 2, \cdots, m, j = 1, 2, \cdots, m+1 \quad (25)$$

Equation (25) implies a misclassification. For a special case, when the following conditions are substituted into (20),

$$c_{ij} = C_i, c_{ik} = 0, i = 1, 2, \cdots, m, j = 1, 2, \cdots, m+1,\ k \neq j \quad (26)$$

one can obtain $I(T,Y) = 0$ or $NH(T|Y) = 1$. These conditions indicate that only one column in C gives the associated class numbers to its every element, and the other columns have zeros for their all elements. □

Equation (20) and two theorems above provide specific properties of classifiers when information-based criteria are used.

Property 1. When an augmented confusion matrix is known after classifications, its associated NI and NH will be uniquely given.

Property 2. When NI or NH is known for a classifier, one is generally unable to determine its associated performance accuracy or other conventional measures.

Property 3. When a classifier shows cases of $A = 1$, but $c_{i(m+1)} = Rej \neq 0$, it also gives $NI = 1$. This property is not desirable for classifications.

For an error analysis of classification problems, we derive the following sensitivity equations of mutual information in binary classifications

$$\frac{\partial I}{\partial TP} = \frac{1}{n}\left[\left(\log_2 \frac{TP}{TP+FP}\right)\mathrm{sgn}(TP) - \left(\log_2 \frac{FN}{FN+TN}\right)\mathrm{sgn}(FN)\right] \quad (27)$$

$$\frac{\partial I}{\partial TN} = \frac{1}{n}\left[-\left(\log_2 \frac{FP}{TP+FP}\right)\mathrm{sgn}(FP) + \left(\log_2 \frac{TN}{FN+TN}\right)\mathrm{sgn}(TN)\right] \quad (28)$$

$$\frac{\partial I}{\partial U_1} = \frac{1}{n}\left[-\left(\log_2 \frac{FN}{FN+TN}\right)\mathrm{sgn}(FN) + \left(\log_2 \frac{U_1}{U_1+U_2}\right)\mathrm{sgn}(U_1)\right] \quad (29)$$

$$\frac{\partial I}{\partial U_2} = \frac{1}{n}\left[-\left(\log_2 \frac{FP}{TP+FP}\right)\mathrm{sgn}(FP) + \left(\log_2 \frac{U_2}{U_1+U_2}\right)\mathrm{sgn}(U_2)\right] \quad (30)$$

where $\mathrm{sgn}(\cdot)$ is a sign function for the reason of $H(0) = 0$.

Property 4. In a generic binary classification, the sensitivity analysis needs four independent equations due to the existing four independent parameters. If no cases for a rejected class, only (27) and (28) are sufficient. The general relations among variables are $TP + FN + U_1 = C_1$, $FP + TN + U_2 = C_2$.

Property 5. Considering a neighbor around the exact solution for a binary classifier, a misclassification on a smaller-number label will produce a bigger change of NI values than on a larger-number label. That is, if $TN < TP$, one will have $\left\|\frac{\partial I}{\partial TN}\right\| > \left\|\frac{\partial I}{\partial TP}\right\|$, or $\left\|\frac{\partial I}{\partial U_2}\right\| > \left\|\frac{\partial I}{\partial U_1}\right\|$.

Property 6. Considering a neighbor around the exact solution for a binary classifier, a misclassification of a sample will produce a bigger change of NI values than assigning that sample into a rejected class. That is, it generally has $\left\|\frac{\partial I}{\partial Z}\right\| > \left\|\frac{\partial I}{\partial U_i}\right\|$, where $Z = TP$ or TN, and $i = 1$ or 2.

Property 7. Considering a binary classification without a rejected class, there exist two maximum points, $NI = 1$, in the 3D plot of "TPR-TNR-NI" (see Fig. 3), which correspond to the cases of either an exact classification or a completely wrong misclassification. The local minimum relations below

$$\frac{TP}{TP+FP} = \frac{FN}{FN+TN},\ \text{or}\ \frac{FP}{TP+FP} = \frac{TN}{FN+TN} \quad (31)$$

will form the bottom curve, with $NI = 0$, in the plot. TPR and TNR are true positive rate (TP/C_1) and true negative rate(TN/C_2), respectively. When these diagonal term variables increase, NI does not show a monotonic relation. Conditions (31) can be further simplified as

$$TP = C_1 - \frac{C_1}{C_2}TN \quad (32)$$

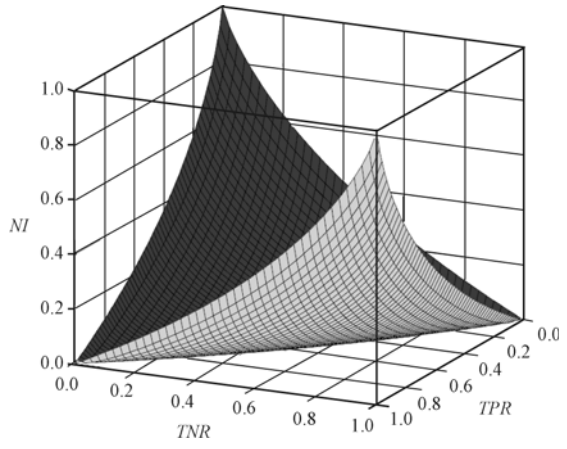

Fig. 3 3D plot of "TPR-TNR-NI" for a binary classification without a rejected class

For a binary classifier, we derived a closed-form solution, $NI = G(A, P, R)$, or the normalized mutual information in relation to the conventional performance indexes, such as accuracy, precision, and recall[24]. In the case without considering a rejected class, this classifier is better to describe its NI with two independent variables. The previous solution showed the nonlinear relationship with respect to three variables, so that one can understand that NI constitutes a suitable measure to balance the conflicting performance criteria in a natural way. In order to transform NI, generally obtained from a single operation point of classifications, into an overall performance index, we suggest the following normalized average mutual information for evaluating classifiers

$$NAI(T,Y) = \frac{1}{S}\sum_Y \sum_T p(t,y)I(T,Y) \quad (33)$$

where S is a scalar for normalizing the index. Since it aver-

ages the mutual information values over the feasible ranges of classifiers in a statistical means, NAI will produce a better evaluation result than with an AUC approach, which applies an evenly averaging scheme.

Theorem 3. The NI criteria defined by Shannon entropy in (1) generally do not show an increasingly monotonic property, whereas the values of the diagonal term variables are increased from reductions of the off-diagonal terms on its confusion matrix.

Proof. Theorem 3 can be proved if one can show NI functions have local minimums with respect to the diagonal term variables of a confusion matrix. For this purpose, we propose a confusion matrix in the following form

$$C = \begin{bmatrix} \cdots & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & c_{i,i} & c_{i,i+1} & \mathbf{0} \\ \mathbf{0} & c_{i+1,i} & c_{i+1,i+1} & \mathbf{0} \\ \cdots & \mathbf{0} & \mathbf{0} & \cdots \end{bmatrix} \tag{34}$$

and impose the local minimums relations below in (34):

$$\begin{aligned} &\frac{c_{i,i}}{c_{i,i}+c_{i+1,i}} = \frac{c_{i,i+1}}{c_{i,i+1}+c_{i+1,i+1}}, \text{ or} \\ &\frac{c_{i+1,i}}{c_{i,i}+c_{i+1,i}} = \frac{c_{i+1,i+1}}{c_{i,i+1}+c_{i+1,i+1}} \end{aligned} \tag{35}$$

From the additivity property of entropy components[3] and Property 7, one can know that the four given variables in (34) and (35) present a zero contribution to NI. In other words, the local minimums exist on NI functions for a confusion matrix described by (34) and (35). Then, NI criteria may not hold a monotonic property with respect to the diagonal term variables, say, $c_{i,i}$ in (34). □

Remark 8. Theorem 3 indicates that Shannon-NI criteria have an intrinsic pitfall of a nonmonotonic property with respect to $c_{i,i}$, which may not prove rational as a measure for evaluating qualities of classifiers. Theorem 3 also demonstrates that the property will also be true when using the joint histogram matrix or joint distribution matrix directly, such as in the studies of image registration[12].

4 Numerical examples

In order to understand well the benefits and limitations of using mutual information criteria in generic classification problems, we present some numerical examples in this section. Table 3 lists ten examples for binary classifications. Among which, M2 ∼ M6 are adopted from [5], where they are denoted as Models A ∼ E, respectively. Although Wallach[27] calculated the mutual information values for M2 ∼ M4 (or Models A ∼ C), the numerical results for M5 and M6 (or Models D and E) were missing. And, it seems that no formula exists in published works to deal with mutual information for classifications including a rejected class.

Table 3 confirms the discussions from [5] about the model selections, if it is evaluated with respect to information criteria. From M7 and M8, one may consider both models are equivalent in performance based on the data of A, P, and R. However, using NI_e data, one can see M8 is a better model than M7. This conclusion is compatible to our intuition in dealing with classification problems, that is, a misclassification from a small number class will pay more loss than a large number one. M9 and M10 show the meaning of Property 3 mentioned in Section 3. Intuitively, any assignment of samples into a rejected class should receive a loss in NI_e, but they do not in M9 and M10. This seems to be a weakness of the current definition of NI. For overcoming this problem, we propose intuitively to modify (20) by the summations of the j-th term from 1 to m, rather than to $m+1$, which shows a better, yet desirable, property for the cases like $c_{i(m+1)} = Rej \neq 0$.

Table 3 Examples in binary classifications including rejection samples ($C_1 = 90$, $C_2 = 10$)

Model	$\begin{bmatrix} TP & FN & U_1 \\ FP & TN & U_2 \end{bmatrix}$	A	P	R	Rej	NI_e
M1	$\begin{bmatrix} 90 & 0 & 0 \\ 0 & 10 & 0 \end{bmatrix}$	1.000	1.000	1.000	0.000	1.000
M2	$\begin{bmatrix} 90 & 0 & 0 \\ 10 & 0 & 0 \end{bmatrix}$	0.900	0.900	1.000	0.000	0.000
M3	$\begin{bmatrix} 80 & 10 & 0 \\ 0 & 10 & 0 \end{bmatrix}$	0.900	1.000	0.889	0.000	0.574
M4	$\begin{bmatrix} 78 & 12 & 0 \\ 0 & 10 & 0 \end{bmatrix}$	0.880	1.000	0.867	0.000	0.534
M5	$\begin{bmatrix} 74 & 6 & 10 \\ 0 & 9 & 1 \end{bmatrix}$	0.933	1.000	0.822	0.110	0.586
M6	$\begin{bmatrix} 78 & 6 & 6 \\ 0 & 5 & 5 \end{bmatrix}$	0.933	1.000	0.867	0.110	0.534
M7	$\begin{bmatrix} 90 & 0 & 0 \\ 1 & 9 & 0 \end{bmatrix}$	0.990	0.989	1.000	0.000	0.831
M8	$\begin{bmatrix} 89 & 1 & 0 \\ 0 & 10 & 0 \end{bmatrix}$	0.990	1.000	0.989	0.000	0.897
M9	$\begin{bmatrix} 90 & 0 & 0 \\ 0 & 9 & 1 \end{bmatrix}$	1.000	1.000	1.000	0.010	1.000
M10	$\begin{bmatrix} 88 & 0 & 2 \\ 0 & 10 & 0 \end{bmatrix}$	1.000	1.000	0.978	0.020	1.000

In Table 4, we present classification examples with three classes. From this table, we can see that mutual information criteria are generally quite effective and efficient in dealing with classification problems for multiple classes, even for a rejected class. M11 shows to be the best in this table even if it has a low level of accuracy ($A = 15\,\%$). Due to $NI_e = 1$, this classifier is able to achieve an exact solution by simple exchanges between two labels. M12 ∼ M16 demonstrate the cases of different classifiers with the same accuracy $A = 95\,\%$. Models 17 and 18 compare the NI_e values for the cases when two samples are either misclassified or assigned into the rejected class. Model 19 shows a case of a confusion matrix having local minimums described by (34) and (35). Although Model 19 has improved classification accuracy ($A = 78\,\%$) based on Model 20 ($A = 68\,\%$), the NI_e does not show such improvement but rather presents an undesirable result. One can obtain the following observations from the given examples:

1) Even for a classifier with multiple classes, a misclassification from a smaller-number label will lead to a bigger change of NI_e values than from a larger-number label.

2) If assigning misclassified samples into a small-number label, its NI_e value of the classifier will receive less impact than the cases of assigning into a large-number label (see M13 vs. M14, and M15 vs. M16).

3) Models 19 and 20 confirm that Shannon-NI criteria may not present a rational measure for assessing qualities of classifiers due to their nonmonotonic properties.

The first observation is consistent with our intuitions

and experiences in selections of classifiers. However, the second one seems to be against the Bayesian principle in classifications, that is, assigning misclassified samples into a large-number label will be safer. The third observation indicates that we need to be cautious when using NI as an evaluation measure, although it can be rational as an optimization function in those studies of image registration or feature selection. More examples can be tested for obtaining useful findings. For convenience, in Appendix, we provide the related program on the open source software Scilab (http://www.scilab.org) so that readers can examine other cases by copying directly from the current document and pasting onto Scilab platform for their uses. When a confusion matrix is given, one can obtain its NI_e and accuracy values easily from the program.

Table 4 Classification examples in three classes

Model	C	A	Rej	NI_e
M11	$\begin{bmatrix} 0 & 0 & 80 & 0 \\ 0 & 15 & 0 & 0 \\ 5 & 0 & 0 & 0 \end{bmatrix}$	0.15	0.00	1.000
M12	$\begin{bmatrix} 75 & 0 & 5 & 0 \\ 0 & 15 & 0 & 0 \\ 0 & 0 & 5 & 0 \end{bmatrix}$	0.95	0.00	0.887
M13	$\begin{bmatrix} 80 & 0 & 0 & 0 \\ 0 & 15 & 0 & 0 \\ 1 & 4 & 0 & 0 \end{bmatrix}$	0.95	0.00	0.753
M14	$\begin{bmatrix} 80 & 0 & 0 & 0 \\ 0 & 15 & 0 & 0 \\ 4 & 1 & 0 & 0 \end{bmatrix}$	0.95	0.00	0.677
M15	$\begin{bmatrix} 80 & 0 & 0 & 0 \\ 1 & 10 & 4 & 0 \\ 0 & 0 & 5 & 0 \end{bmatrix}$	0.95	0.00	0.811
M16	$\begin{bmatrix} 80 & 0 & 0 & 0 \\ 4 & 10 & 1 & 0 \\ 0 & 0 & 5 & 0 \end{bmatrix}$	0.95	0.00	0.693
M17	$\begin{bmatrix} 79 & 0 & 1 & 0 \\ 0 & 14 & 1 & 0 \\ 0 & 0 & 5 & 0 \end{bmatrix}$	0.98	0.00	0.909
M18	$\begin{bmatrix} 79 & 0 & 0 & 1 \\ 0 & 14 & 0 & 1 \\ 0 & 0 & 5 & 0 \end{bmatrix}$	1.00	0.02	0.977
M19	$\begin{bmatrix} 50 & 0 & 0 & 0 \\ 0 & 24 & 16 & 0 \\ 0 & 6 & 4 & 0 \end{bmatrix}$	0.78	0.00	0.735
M20	$\begin{bmatrix} 50 & 0 & 0 & 0 \\ 0 & 14 & 26 & 0 \\ 0 & 6 & 4 & 0 \end{bmatrix}$	0.68	0.00	0.746

5 Conclusions

In machine learning, evaluation criteria or learning targets can vary due to the study viewpoints or application background[28]. In classifier designs, a selection of evaluation criteria may influence the performance of the associated classifiers directly. For example, the least-mean-squared criteria (or LME) may give a misclassification result even for linearly separable problems[25]. However, after a minor change to the same learning criteria, one is able to achieve an exact solution to the same problems[29]. Up to now, a selection of evaluation criteria has remained an open issue in the theoretical studies of machine learning. Although many investigations are reported in related subjects, a systematic study is needed for dealing with classification problems. Information theory will lead to a new direction in machine learning but relations of information entropy to other measures, such as knowledge granularity[30] and correlation coefficient[31], are fundamental subjects in the studies. As a preliminary study, this work explored the relations between mutual information and the conventional performance measures. We derived theoretical formulas and interpretations to the generic classifications, which may include a rejected class. From the results, simple interpretations to the use of mutual information criteria can be stated below:

The principle behind machine learning of using entropy-type criteria is to transform disordered data sets into ordered data ones. For classification problems, this transformation is made on label data sets, which is different from that on feature data sets for clustering problems.

More detailed descriptions about the advantages of using Shannon-information-based criteria in classification problems can be summarized as follows:

1) Entropy-type criteria provide classifier designers with unique information, which will significantly enlarge the searching range for the potential classifiers, which may be neglected by using the conventional performance criteria.

2) Entropy-type criteria present a simple and generic framework of dealing with higher-order stochastic variables or processes in various applications, including classifications.

3) Entropy-type criteria produce a single and objective index for balancing the conflicting performance measures naturally and globally in classification problems, even in the cases of assigning samples into a rejected class.

4) The computational complexity of entropy-type criteria is reasonably low for label data, that is, O(m^2). Therefore, these criteria are suitable for classification of problems.

However, limitations of Shannon-entropy-type criteria are also observed as:

1) The uncertainty concept from entropy-type criteria is not a common concern or requirement from most classifier designers and users.

2) Shannon-entropy-type criteria do not hold monotonic properties, nor the one-to-one correspondence, to the conventional performance measures.

3) To reach a reasonable evaluation of classifiers, one still needs to use calculations from the conventional performance measures as its assistants.

Appendix. Scilab code used in Tables 3 and 4

```
// Scilab code for calculating normalized mutual information
// from a given m-by-(m + 1) confusion matrix.
c=[79 0 0 1
    0 14 0 1
    0  0 5 0];
n=sum(c);              // = number of total samples
m=length(c(:,1));      // = number of exact classes
Ci=sum(c,'c');         // = numbers of exact labels
Cp=sum(c,'r');         // = numbers of prediction labels
NI_num=0;              // = numerator of NI in (20)
NI_den=0;              // = denominator of NI in (20)
for i=1 : m
```

```
  NI_den=NI_den+Ci(i)*log2(Ci(i)/n);
  for j=1 : m+1
    if c(i,j) > 0 then
      if Ci(i)*Cp(j) > 0 then
        NI_num=NI_num+c(i,j)*log2(c(i,j)/Ci(i)/(Cp(j)/n));
      end
    end
  end
end
NI=-NI_num/NI_den;  // Normalized Mutual Information
if sum(c(1:m,1:m)) > 0 then
  A=sum(diag(c))/sum(c(1:m,1:m))      // Accuracy
else
  A=0;
end
if length(c(:,1)) < 3 then                  // Binary classifier
  if Cp(1)>0 then
    P=c(1,1)/Cp(1)                          // Precision
  else
    P=0;
  end
  R=c(1,1)/Ci(1)                            // Recall
end
Rej=sum(c(:,m+1))/n                         // Rejection Rate
```

References

1 Haykin S. *Neural Networks: A Comprehensive Foundation (Second Edition)*. New York: Printice Hall, 1999

2 Xu D X. Energy, Entropy and Information Potential for Neural Computation [Ph. D. dissertation], University of Florida, 1998

3 Cover T M, Thomas J A. *Elements of Information Theory*. New York: John Wiley, 1991

4 Mitchell T M. *Machine Learning*. New York: McGraw-Hill, 1997

5 Mackay D J C. *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press, 2003. 140, 532−533

6 Linsker R. Self-organization in a perceptual network. *Computer*, 1988, **21**(3): 105−117

7 Battiti R. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 1994, **5**(4): 537−550

8 Kwak N, Choi C H. Input feature selection for classification problems. *IEEE Transactions on Neural Networks*, 2002, **13**(1): 143−159

9 Peng H C, Long F H, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, **27**(8): 1226−1238

10 Huang Jin-Jie, Lv Ning, Li Shuang-Quan, Cai Yun-Ze. Feature selection for classificatory analysis based on information-theoretic criteria. *Acta Automatica Sinica*, 2008, **34**(3): 383−392

11 Comon P. Independent component analysis: a new concept? *Signal Processing*, 1994, **36**(3): 287−314

12 Maes F, Collignon A, Vandermeulen D, Marchal G, Suetens P. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 1997, **16**(2): 187−198

13 Tang Min. A novel image registration method combining morphological gradient mutual information with multiresolution optimizer. *Acta Automatica Sinica*, 2008, **34**(3): 246−250 (in Chinese)

14 Principe J C, Xu D X, Zhao Q, Fisher J W. Learning from examples with information-theoretic criteria. *Journal of VLSI Signal Processing Systems*, 2000, **26**(1-2): 61−77

15 Sridhar D V, Bartlett E B, Seagrave R C. Information theoretic subset selection for neural network models. *Computers and Chemical Engineering*, 1998, **22**(4-5): 613−626

16 Press W H, Flannery B P, Teukolsky S A, Vetterling W T. *Numerical Recipes in C*. Cambridge: Cambridge University Press, 1988

17 Strehl A, Ghosh J. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 2003, **3**: 583−617

18 Quinlan J R. Introduction of decision trees. *Machine Learning*, 1986, **1**(1): 81−106

19 Quinlan J R. *C4.5: Programs for Machine Learning*. San Francisco: Morgan Kaufmann, 1993

20 Ding Xiao-Qing, Wu You-Shou. Information entropy theory in pattern recognition. *Acta Electronica Sinica*, 1993, **21**(8): 1−8 (in Chinese)

21 Fano R M. *Transmission of Information: A Statistical Theory of Communication*. New York: MIT, 1961

22 Hellman M, Raviv J. Probability of error, equivocation, and the Chernoff bound. *IEEE Transactions on Information Theory*, 1970, **16**(4): 368−372

23 Eriksson T, Kim S, Kang H G, Lee C. An information-theoretic perspective on feature selection in speaker recognition. *IEEE Signal Processing Letters*, 2005, **12**(7): 500−503

24 Wang Yong, Hu Bao-Gang. A study on integrated evaluating kernel classification performance using statistical methods. *Chinese Journal of Computers*, 2008, **31**(6): 64−74 (in Chinese)

25 Duda R O, Hart P E, Stork D G. *Pattern Classification (Second Edition)*. New York: John Wiley, 2001

26 Wagner R F, Metz C E, Campbell G. Assessment of medical imaging systems and computer aids: a tutorial review. *Academic Radiology*, 2007, **14**(6): 723−748

27 Wallach H. Evaluation metrics for hard classifiers [Online], available: http://www.inference.phy.cam.ac.uk/hmw26/, November 3, 2007

28 Hu Bao-Gang, Wang Yong, Yang Shuang-Hong, Qu Han-Bing. How to add transparency to artificial neural networks? *Pattern Recognition and Artificial Intelligence*, 2007, **20**(1): 72−84 (in Chinese)

29 Yang S H, Hu B G. A stagewise least square loss function for classification. In: Proceedings of the 2008 SIAM International Conference on Data Mining. Japan: IEEE, 2008. 120−131

30 Miao Dou-Qian, Wang Jue. On the relationships between information entropy and roughness of knowledge in rough set theory. *Pattern Recognition and Artificial Intelligence*, 1998, **11**(1): 34−40 (in Chinese)

31 Borga M. Learning Multidimensional Signal Processing [Ph. D. dissertation], Linkoping University, 1998

HU Bao-Gang Received his master degree from University of Science and Technology Beijing in 1983, and Ph. D. degree from McMaster University, Canada in 1993, all in Mechanical Engineering. Currently, he is a professor at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interest covers intelligent systems, pattern recognition, and plant growth modeling. Corresponding author of this paper. E-mail: hubg@nlpr.ia.ac.cn

WANG Yong Received his Ph. D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences in 2008. He is currently a postdoctor at Graduate University of Chinese Academy of Sciences. His research interest covers pattern recognition, knowledge discovery, data mining, and extensible business reporting language. E-mail: wangyong@gucas.ac.cn