What Are the Differences Between Bayesian Classifiers and Mutual-Information Classifiers?

Bao-Gang Hu, *Senior Member, IEEE*

Abstract—**In this paper, both Bayesian and mutual-information classifiers are examined for binary classifications with or without a reject option. The general decision rules are derived for Bayesian classifiers with distinctions on error types and reject types. A formal analysis is conducted to reveal the parameter redundancy of cost terms when abstaining classifications are enforced. The redundancy implies an intrinsic problem of nonconsistency for interpreting cost terms. If no data are given to the cost terms, we demonstrate the weakness of Bayesian classifiers in class-imbalanced classifications. On the contrary, mutual-information classifiers are able to provide an objective solution from the given data, which shows a reasonable balance among error types and reject types. Numerical examples of using two types of classifiers are given for confirming the differences, including the extremely class-imbalanced cases. Finally, we briefly summarize the Bayesian and mutual-information classifiers in terms of their application advantages and disadvantages, respectively.**

Index Terms—**Abstaining classifier, Bayes, cost-sensitive learning, entropy, error types, mutual information, reject types.**

I. Introduction

BAYESIAN principle provides a powerful and formal means of dealing with statistical inference in data processing, such as classifications [1], [2]. If the classifiers are designed based on this principle, they are called Bayesian classifiers in this paper. In recent years, cost-sensitive learning and class-imbalanced learning have received much attention in classifications [3]–[6]. Within the imbalanced, or skewed, data sets, "the ratio of the small to the large classes can be drastic such as 1–100, 1–1000, or 1–10 000 (and sometimes even more)" [7]. In fact, the related subjects are not a new challenge but a more crucial concern than before for increasing the needs of searching useful information in big data processing. Binary classifications will be a basic problem in such application background. Classifications based on cost terms for the tradeoff of error types are a conventional subject in medical diagnosis. Misclassification from type I error (or false positive) or from type II error (or false negative) is significantly different in the context of medical practices. Therefore, cost terms play a key role in class-imbalanced learning [6], [8].

It was recognized that Chow [9] was "among the earliest to use Bayesian decision theory for pattern recognition" [2]. In [10], Chow first derived the error-reject tradeoff formulas, but assumed no distinctions among errors and rejects. The more general settings for distinguishing error types and reject types were reported in [11]–[15], but they generally require cost terms. To overcome the problems of presetting cost terms manually, Pietraszek [14] proposed two learning models, such as bounded-abstention and bounded-improvement, which are usually related to the performance constraints [16]. If constraints are given by total amount of either errors or rejects, they may result in no distinctions among error or reject types.

In addition to a kind of ambiguity reject studied in [10], the other kind of distance reject was also considered in [17]. Ambiguity reject is made to a pattern located in an ambiguous region between/among classes. Distance reject represents a pattern, which is conventionally called an outlier in statistics [2]. Ha [18] proposed another important kind of reject, called class-selective reject, which defines a subset of classes. This scheme is more suitable to multiple-class classifications. For example, in three-class problems, Ha's classifiers will output the predictions including ambiguity reject between Classes 1 and 2, ambiguity reject among Classes 1, 2, and 3, and the other rejects from class combinations. Multiple rejects with such distinctions will be more informative than a single ambiguity reject. Among all these investigations, the Bayesian principle is applied again for their design guideline of classifiers.

While the Bayesian inference principle is widely applied in classifications [1], [2], [19], [20], another principle based on the mutual-information concept is rarely adopted for designing classifiers. Mutual information is one of the important definitions in entropy theory [21]. Entropy is considered as a measure of uncertainty within random variables, and mutual information describes the relative entropy between two random variables [19]. If classifiers seek to maximize the relative entropy for their learning target, we refer them to mutual-information classifiers. It seems that Quinlan [22] was among the earliest to apply the concept of mutual information (but called information gain in his famous iterative dichotomiser 3 algorithm) in constructing the decision tree. Kvålseth [23] introduced the definition of normalized mutual information (NI) for assessing a contingency table, which laid down the foundation on the relationship between a confusion matrix and mutual information. Being pioneers in using an information-based criterion for classifier evaluations,

Manuscript received March 12, 2012; accepted July 20, 2013. Date of publication August 15, 2013; date of current version January 10, 2014. This work was supported by the NSFC under Grant 61075051 and Grant 61273196.

The author is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: hubg@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2013.2274799

Kononenko and Bratko [24] suggested the term information score, which was equivalent to the definition of mutual information. A research team led by Principe [25], [26] proposed a general framework, called information-theoretic learning, for designing various learning machines, in which he suggested that mutual information, or other information-theoretic criteria, can be set as an objective function in classifier learning. Mackay [19] suggested to apply mutual information for ranking classification results. Wang and Hu [27] derived the nonlinear relations between mutual information and the conventional performance measures, such as accuracy, precision, recall, and F1 measure for binary classifications. In [28] and [29], information-theoretic measures were studied on classification evaluations for binary and multiple-class problems with/without a reject option. The advantages and limitations of mutual-information measures were discussed in [28]. No systematic investigation is, however, reported for a theoretical comparison between Bayesian and mutual-information classifiers in the literature.

For comparing with Bayesian classifiers, this paper derives much from and therefore extends to Chow's work [10] by distinguishing error types and reject types. To achieve analytical tractability without losing the generality, a strategy of adopting the simplest yet most meaningful assumptions to classification problems is pursued for investigations. The following assumptions are given in the same way as those in the closed-form studies of Bayesian classifiers in [10] and [2].

A1. Classifications are made for two categories (or classes) over the feature variables.

A2. All probability distributions of feature variables are exactly known.

We may argue that the assumptions above are extremely restricted to offer practical generality in solving real-world problems. In fact, the power of Bayesian classifiers does not stay within their exact solutions to the theoretical problems, but appear from their generic inference principle in guiding real applications, even in the extreme approximations to the theory. We fully recognize that the assumption of complete knowledge on the relevant probability distributions is never the case in real-world problems [30]. The closed-form solutions of Bayesian classifiers on binary classifications in [2] and [10] have demonstrated the useful design guidelines that are applicable to multiple classes [18]. The author believes that the analysis based on the above-mentioned assumptions will provide sufficient information for revealing the basic differences between Bayesian and mutual-information classifiers, while the intended simplifications will benefit readers to reach a better, or deeper, understanding to the advantages and limitations of each type of classifiers.

This paper proposes mutual-information classifiers and investigates the differences between these classifiers and Bayesian classifiers, especially for the settings with a reject option. This paper mainly contributes in the following aspects.

1) Theoretical basis of Bayesian classifiers is further explored. Three novel theorems are derived in the following for binary classifications: a) general Bayesian rules for distinguishing error types and reject types; b) parameter redundancy to cost terms for abstaining classifications; and c) the Bayesian error behavior and bound in class imbalanced problems.
2) The unique features of mutual-information classifiers are discovered, for which Bayesian classifiers do not possess. The former type of classifiers does not require the cost terms as input data in class imbalanced learning, is capable of automatically balancing error types and reject types from a given data set, and provides an interpretation to the learning rule of less costs more in classifications.

The rest of this paper is organized as follows. Section II presents a general decision rule of Bayesian classifiers with or without a reject option. Section III provides the basic formulas for mutual-information classifiers. Section IV conducts the comparisons between two types of classifiers, and numerical examples are given to highlight the distinct features in their applications. The question presented in the title of this paper is concluded by a simple answer in Section V.

II. Bayesian Classifiers With a Reject Option

A. General Decision Rule for Bayesian Classifiers

Let $\mathbf{x}$ be a random pattern satisfying $\mathbf{x} \in \mathbf{X} \subset R^d$, which is in a d-dimensional feature space and will be classified. The true (or target) state t of $\mathbf{x}$ is within the finite set of two classes, $t \in T = \{t_1, t_2\}$, and the possible decision output $y = f(\mathbf{x})$ is within three classes, $y \in Y = \{y_1, y_2, y_3\}$, where f is a function for classifications and y_3 is a reject class. Let $p(t_i)$ be the prior probability of class t_i and $p(\mathbf{x}|t_i)$ be the conditional probability density function of $\mathbf{x}$ given that it belongs to class t_i. The posterior probability $p(t_i|\mathbf{x})$ is calculated through the Bayes formula [2]

$$p(t_i|\mathbf{x}) = \frac{p(\mathbf{x}|t_i)p(t_i)}{p(\mathbf{x})} \tag{1}$$

where $p(\mathbf{x})$ is the mixture density for normalizing the probability. With the posterior probability, the Bayesian rule assigns a pattern $\mathbf{x}$ into the class that has the highest posterior probability. Chow [10] investigated rejects under the Bayesian principle for the first time. The purpose of the reject rule is to minimize the total risk (or cost) in classifications. Suppose λ_{ij} is a cost term for the true class of a pattern to be t_i, but decided as y_j. Then, the conditional risk for classifying a particular $\mathbf{x}$ into y_j is defined as follows:

$$\mathrm{Risk}(y_j|\mathbf{x}) = \sum_{i=1}^{2} \lambda_{ij} p(t_i|\mathbf{x}) = \sum_{i=1}^{2} \lambda_{ij} \frac{p(\mathbf{x}|t_i)p(t_i)}{p(\mathbf{x})}$$
$$j = 1, 2, 3. \tag{2}$$

Note that the definition of λ_{ij} in this paper is a bit different from that in [2], so that λ_{ij} will form a 2×3 cost matrix. Chow [10] assumed the initial constraints on λ_{ij} from the intuition in classifications

$$\lambda_{ik} > \lambda_{i3} > \lambda_{ii} \geq 0, \quad i \neq k, \quad i = 1, 2, \quad k = 1, 2. \tag{3}$$

The constraints imply that, within the same class, a misclassification will suffer a higher cost than a rejection, and a rejection will cost more than a correct classification. The total risk for the decision output y will be [2]

$$\mathrm{Risk}(y) = \int_V \sum_{j=1}^{3}\sum_{i=1}^{2} \lambda_{ij}\, p(t_i|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \tag{4}$$

with integration over the entire observation space V.

Definition 1 (Bayesian Classifier): Given (2), if a classifier is formed from the minimization of its risk over all patterns

$$y^* = \arg\min_y \mathrm{Risk}(y) \tag{5a}$$

or on a given pattern $\mathbf{x}$

$$\text{Decide } \; y_j \;\text{ if } \; \mathrm{Risk}(y_j|\mathbf{x}) = \min_i \mathrm{Risk}(y_i|\mathbf{x}) \tag{5b}$$

this classifier is called Bayesian classifier, or Chow's abstaining classifier [13]. The term of $\mathrm{Risk}(y^*)$ is usually called Bayesian risk, or Bayesian error in the cases that zero–one cost terms ($\lambda_{11} = \lambda_{22} = 0, \lambda_{12} = \lambda_{21} = 1$) are used for no rejection classifications [2].

In [10], a single threshold for a reject option was investigated. This setting was obtained from the assumption that cost terms are applied without distinction among the errors and rejects. Following Chow's approach [10] but with extension to the general cases in cost terms, we can derive the general decision rule on the rejection for Bayesian classifiers.

Theorem 1: The general decision rule for Bayesian classifiers are as follows:

$$\begin{aligned}
&\text{Decide } y_1 \text{ if } \frac{p(\mathbf{x}|t_1)p(t_1)}{p(\mathbf{x}|t_2)p(t_2)} > \delta_1\\
&\text{No rejection: } \delta_1 = \frac{\lambda_{21}-\lambda_{22}}{\lambda_{12}-\lambda_{11}}\\
&\text{Rejection: } \delta_1 = \frac{\lambda_{21}-\lambda_{23}}{\lambda_{13}-\lambda_{11}}
\end{aligned} \tag{6a}$$

$$\begin{aligned}
&\text{Decide } y_2 \text{ if } \frac{p(\mathbf{x}|t_1)p(t_1)}{p(\mathbf{x}|t_2)p(t_2)} \le \delta_2\\
&\text{No rejection: } \delta_2 = \frac{\lambda_{21}-\lambda_{22}}{\lambda_{12}-\lambda_{11}}\\
&\text{Rejection: } \delta_2 = \frac{\lambda_{23}-\lambda_{22}}{\lambda_{12}-\lambda_{13}}
\end{aligned} \tag{6b}$$

$$\begin{aligned}
&\text{Decide } y_3 \text{ if } \frac{T_{r2}}{1-T_{r2}} = \frac{\lambda_{23}-\lambda_{22}}{\lambda_{12}-\lambda_{13}}\\
&< \frac{p(\mathbf{x}|t_1)p(t_1)}{p(\mathbf{x}|t_2)p(t_2)} \le \frac{\lambda_{21}-\lambda_{23}}{\lambda_{13}-\lambda_{11}} = \frac{1-T_{r1}}{T_{r1}}
\end{aligned} \tag{6c}$$

$$\begin{aligned}
&\text{Subject to } 0 < \frac{\lambda_{23}-\lambda_{22}}{\lambda_{12}-\lambda_{13}} < \frac{\lambda_{21}-\lambda_{22}}{\lambda_{12}-\lambda_{11}}\\
&< \frac{\lambda_{21}-\lambda_{23}}{\lambda_{13}-\lambda_{11}}
\end{aligned} \tag{6d}$$

$$\begin{aligned}
&\text{No rejection: } T_{r1} = T_{r2} = 0.5\\
&\text{Rejection: } 0 < T_{r1} + T_{r2} \le 1.
\end{aligned} \tag{6e}$$

Equation (6c) applies the definition of two thresholds (called rejection thresholds in [10]), T_{r1} and T_{r2}.

Proof: See Appendix A. ■

Note that (6d) suggests general constraints over λ_{ij}. The necessity for having such constraints is explained in Appendix A. A graphical interpretation to the two thresholds is shown in Fig. 1. With (6c), the thresholds can be calculated from the following formulas:

$$\begin{aligned}
T_{r1} &= \frac{\lambda_{13}-\lambda_{11}}{\lambda_{13}-\lambda_{11}+\lambda_{21}-\lambda_{23}}\\
T_{r2} &= \frac{\lambda_{23}-\lambda_{22}}{\lambda_{12}-\lambda_{13}+\lambda_{23}-\lambda_{22}}.
\end{aligned} \tag{7}$$

Equation (7) describes general relations between thresholds and cost terms in binary classifications, which enable the classifiers to make the distinctions among errors and rejects. The special settings of Chow's rules [10], $T_r = \lambda_{13} = \lambda_{23}$ for $\lambda_{11} = \lambda_{22} = 0$, $\lambda_{12} = \lambda_{21} = 1$, can be derived from (7). The studies in [11], [14], and [15] generalized Chow's rules [10] by distinguishing error and reject types for the relations of cost terms. Their constraint relations of missing the terms λ_{11} and λ_{22} are not theoretically general, yet sufficient for applications. Up to now, it seems no one has reported the general constraints, (6d), in the literature.

By applying (1) and the constraint $p(t_1|\mathbf{x}) + p(t_2|\mathbf{x}) = 1$, we can achieve the decision rules from (6) with respect to the posterior probabilities and thresholds in a simple and better form for abstaining classifiers

$$\begin{aligned}
&\text{Decide } y_1, \text{ if } p(t_1|\mathbf{x}) > 1 - T_{r1}\\
&\text{Decide } y_2, \text{ if } p(t_2|\mathbf{x}) \ge 1 - T_{r2}\\
&\text{Decide } y_3, \text{ otherwise}\\
&\text{Subject to } 0 < T_{r1} + T_{r2} \le 1.
\end{aligned} \tag{8}$$

In comparison with the decision rules of (6), which are expressed in terms of the likelihood ratio, (8) together with Fig. 1 shows a better view for users to understand abstaining Bayesian classifiers. A plot of posterior probabilities shows advantages over a plot of the likelihood ratio (Fig. 2.3, in [2]) for determining rejection thresholds. Note that in Fig. 1, the plots are depicted on a 1-D variable for Gaussian distributions of X. The simplification supports the suggestions by Duda *et al.* that we "should not obscure the central points illustrated in our simple example" [2]. Two sets of geometric points are shown for the plots. One set is called crossover points, denoted by x_{ci}, which are formed from two curves of $p(t_1|x)$ and $p(t_2|x)$. In addition, the other is termed as boundary points, denoted by x_{bj}. The boundary points partition classification regions in 1-D problems. In a no-rejection case, the boundary points are controlled by the ratio of $(\lambda_{21} - \lambda_{22})/(\lambda_{12} - \lambda_{11})$. In abstaining classifications, those points are determined from two thresholds, respectively. If a multiple-dimensional problem is considered, the analysis with a reject option will become significantly tedious. For example, six types of boundaries are formed even for classifications in two dimensions (Fig. 2.14, in [2]).

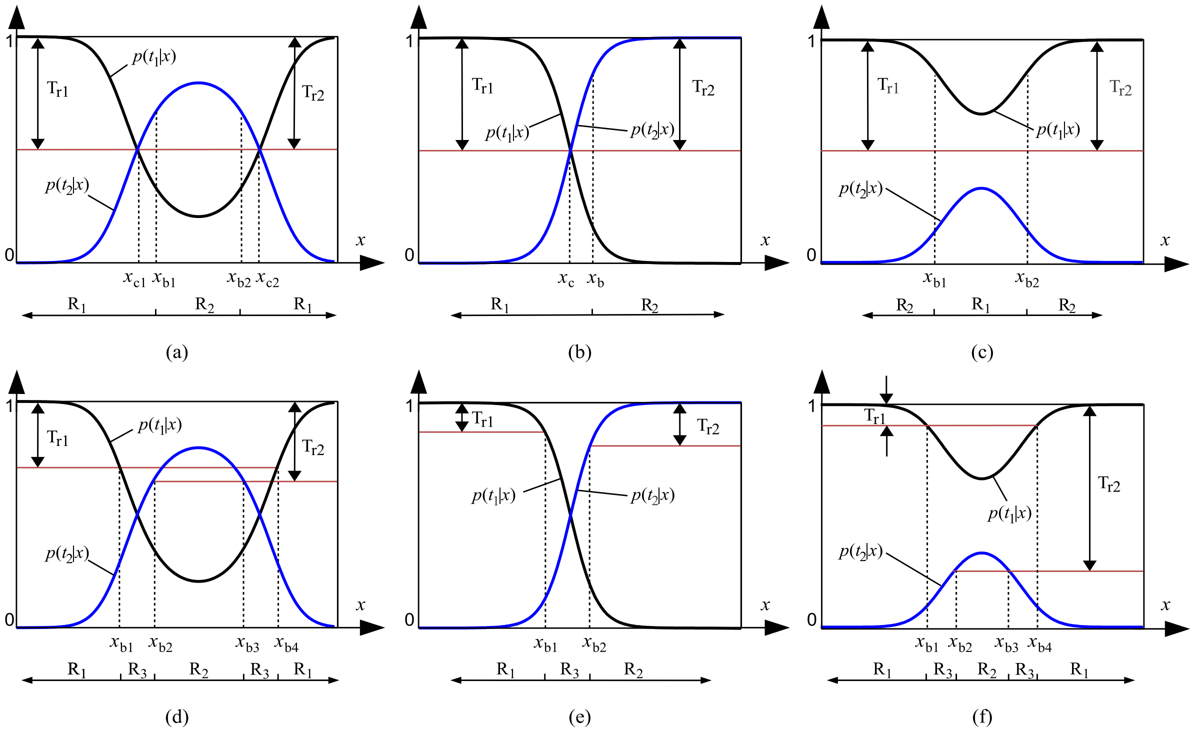


Fig. 1. Rejection scenarios from the plots of $p(t_i|x)$ for univariate Gaussian distributions. x_{ci} are the crossover points. x_{bj} are the boundary points. R_1–R_3 are the classification regions of Class 1, Class 2, and the reject class, respectively. T_{r1} and T_{r2} are the rejection thresholds of Classes 1 and 2, respectively.

With the exact knowledge of $p(t_i)$, $p(\mathbf{x}|t_i)$, and λ_{ij}, we can calculate Bayesian risk from the following equation:

$$\begin{aligned}\text{Risk}(y^*) &= \lambda_{11}CR_1 + \lambda_{12}E_1 + \lambda_{13}\text{Rej}_1 + \lambda_{22}CR_2 \\ &\quad + \lambda_{21}E_2 + \lambda_{23}\text{Rej}_2 \\ &= \lambda_{11}\int_{R_1} p(t_1)p(\mathbf{x}|t_1)d\mathbf{x} + \lambda_{12}\int_{R_2} p(t_1)p(\mathbf{x}|t_1)d\mathbf{x} \\ &\quad + \lambda_{13}\int_{R_3} p(t_1)p(\mathbf{x}|t_1)d\mathbf{x} + \lambda_{21}\int_{R_1} p(t_2)p(\mathbf{x}|t_2)d\mathbf{x} \\ &\quad + \lambda_{22}\int_{R_2} p(t_2)p(\mathbf{x}|t_2)d\mathbf{x} + \lambda_{23}\int_{R_3} p(t_2)p(\mathbf{x}|t_2)d\mathbf{x}\end{aligned} \tag{9}$$

where CR_i, E_i, and Rej_i are the probabilities of correct recognition, error, and rejection for the ith class in the classifications, respectively, and R_1–R_3 are the classification regions of Class 1, Class 2, and the reject class, respectively. The general relations among CR_i, E_i, and Rej_i for binary classifications are given by [10]

$$\begin{aligned}&CR_1+CR_2+E_1+E_2+\text{Rej}_1+\text{Rej}_2 = CR+E+\text{Rej} = 1 \\ &\text{and} \quad A = \frac{CR}{CR+E}\end{aligned} \tag{10}$$

where CR, E, and Rej are the total correct recognition, total error, and total reject rates, respectively, and A is the accuracy rate of classifications.

B. Parameter Redundancy Analysis of Cost Terms

Bayesian classifiers present one of the general tools for cost-sensitive learning. From this perspective, there exists a need for a systematic investigation into a parameter redundancy analysis of cost terms for Bayesian classifiers, which appears missing for a reject option. This section will attempt to develop a theoretical analysis of parameter redundancy for cost terms.

In Bayesian classifiers, when all cost terms are given along with the other relevant knowledge about classes, a unique set of solutions will be obtained. This phenomenon, however, does not show that all cost terms will be independent for determining the final results of Bayesian classifiers. In the following, a parameter redundancy analysis is conducted because it suggests a theoretical basis for a better understanding of relations among the cost terms and the outputs of Bayesian classifiers. Different from the functions to be known in the analysis [31], [32], we derive a theorem from the functionals in (4) and (5) so that it holds generality for any distributions of features. Let a parameter vector be defined as $\theta = \{\theta_1, \theta_2, \ldots, \theta_p\} \in \mathbf{S}$, where p is the total number of parameters in a model $f(\mathbf{x}, \theta)$ and $\mathbf{S}$ is the parameter space.

Definition 2 (Parameter Redundancy [31]): A model $f(\mathbf{x}, \theta)$ is considered to be parameter redundant if it can be expressed in terms of a smaller parameter vector $\beta = \{\beta_1, \beta_2, \ldots, \beta_q\} \in \mathbf{S}$, where $q < p$.

Definition 3 (Independent Parameters): A model $f(\mathbf{x}, \theta)$ is said to be governed by independent parameters if it can be expressed in terms of the smallest size of parameter vector $\beta = \{\beta_1, \beta_2, \ldots, \beta_m\} \in \mathbf{S}$, where $m \leq p$. Let $N_{IP}(\theta)$ denote the total number of independent parameters for the model $f(\mathbf{x}, \theta)$, which is equal to m.

Definition 4 (Parameter Function, Input Parameters, and Intermediate Parameters): Suppose two sets of parameter vectors are denoted by $\theta = \{\theta_1, \theta_2, \ldots, \theta_p\} \in \mathbf{S}_1$, and $\gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_q\} \in \mathbf{S}_2$. If for a model there exists

$f(\mathbf{x}, \theta) = f(\mathbf{x}, \psi(\theta))$ for $\psi(\theta) = \gamma\colon \mathbf{S}_1 \to \mathbf{S}_2$, we call ψ a parameter function, θ_i are the input parameters and γ_j are the intermediate parameters.

Lemma 1: Suppose a model holds the relation $f(\mathbf{x}, \theta) = f(\mathbf{x}, \psi(\theta))$ from Definition 4. The total number of independent parameters of θ, denoted as $N_{IP}(f, \theta)$ for the model f will be given in a form of

$$N_{\text{IP}}(f, \theta) \leq \min(p, q) \tag{11}$$

where the symbol min denotes a minimum operation.

Proof: Suppose $f(\mathbf{x}, \theta = \{\theta_1, \theta_2, \ldots, \theta_p\})$ without a parameter function, we can prove that $N_{IP}(f, \theta) \leq \min(p)$. According to Definition 2, any increase of its size of θ over p will produce a parameter redundancy in the model. Definition 3 shows that the vector size p will be an upper bound (UB) for $N_{IP}(f, \theta)$ in this situation. In the same principle, given $f(\mathbf{x}, \theta) = f(\mathbf{x}, \psi(\theta))$ and $\gamma = \psi(\theta)$, the lowest parameter size within θ and γ, will be the UB of $f(\mathbf{x}, \theta)$. ■

For Bayesian classifiers defined by (5a) and (6), we can rewrite it in the following form:

$$y^* = \arg\min \text{Risk}(y, \{\theta_\lambda, \theta_{\mathbf{C}}\}) \tag{12}$$

where $\theta_\lambda = (\lambda_{11}, \lambda_{12}, \lambda_{13}, \lambda_{21}, \lambda_{22}, \lambda_{23})$ and $\theta_{\mathbf{C}} = (p(t_1), p(t_2), p(\mathbf{x}|t_1), p(\mathbf{x}|t_2))$ in binary classifications. Two disjoint sets of parameters, $\theta_\lambda \cap \theta_{\mathbf{C}} = \emptyset$, are given so that one is able to conduct an individual analysis of θ_λ. Similarly, from (8), the abstaining Bayesian classifiers will be expressed by

$$y^* = \arg\min \text{Risk}(y, \{\theta_{\mathbf{Tr}}, \theta_{\mathbf{C}}\}) \tag{13}$$

where $\theta_{\mathbf{Tr}} = (T_{r1}, T_{r2})$. Their total error and reject will be unchanged using either set of parameters

$$\begin{aligned} E(y^*, \{\theta_\lambda, \theta_{\mathbf{C}}\}) &= E(y^*, \{\theta_{\mathbf{Tr}}, \theta_{\mathbf{C}}\}) \\ \text{Rej}(y^*, \{\theta_\lambda, \theta_{\mathbf{C}}\}) &= \text{Rej}(y^*, \{\theta_{\mathbf{Tr}}, \theta_{\mathbf{C}}\}) \end{aligned} \tag{14}$$

where θ_λ are usually input parameters, but $\theta_{\mathbf{Tr}}$ can serve as either intermediate parameters in (6) or input ones in (8).

Theorem 2: In abstaining binary classifications, the total number of independent parameters within the cost terms for defining Bayesian classifiers should be at most two, that is, $N_{IP}(f, \theta_\lambda) \leq 2$. Therefore, applications of cost terms of $\theta_\lambda = (\lambda_{11}, \lambda_{12}, \lambda_{13}, \lambda_{21}, \lambda_{22}, \lambda_{23})$ in the traditional cost-sensitive learning will exhibit a parameter redundancy for calculating Bayesian $E(y^*)$ and $\text{Rej}(y^*)$ even after assuming $\lambda_{11} = \lambda_{22} = 0$, and $\lambda_{12} = 1$ as the conventional way in classifications [8], [13].

Proof: For the given $\theta_{\mathbf{C}}$, a decision function f is determined by $\theta_{\mathbf{Tr}}$. Equation (7) describes a parameter function between two sets of parameters so that a relation of $\theta_{\mathbf{Tr}} = \psi(\theta_\lambda)$ holds in abstaining classifications. According to Lemma 1, we can have $N_{IP}(f, \theta_\lambda) \leq \min(p = 6, q = 2) = 2$, where p and q are the parameter sizes of θ_λ and $\theta_{\mathbf{Tr}}$, respectively. However, when imposing three constraints on $\lambda_{11} = \lambda_{22} = 0$, and $\lambda_{12} = 1$, θ_λ will provide three free parameters in the cost matrix in the following form:

$$\begin{aligned} \lambda_{21} &= \lambda_{21} \\ \lambda_{13} &= \frac{T_{r1}(T_{r2} * \lambda_{21} + T_{r2} - \lambda_{21})}{T_{r1} + T_{r2} - 1} \\ \lambda_{23} &= \frac{T_{r2}(T_{r1} * \lambda_{21} + T_{r1} - 1)}{T_{r1} + T_{r2} - 1} \end{aligned} \tag{15}$$

which implies a parameter redundancy for calculating Bayesian $E(y^*)$ and $\text{Rej}(y^*)$. ■

Remark 1: Theorem 2 describes that Bayesian classifiers with a reject option will suffer a difficulty of uniquely interpreting cost terms. For example, we can even enforce the following two settings:

$$\begin{cases} \lambda_{11} = 0, \ \lambda_{12} = 1, \ 0 \leq \lambda_{13} \leq 1 \\ \lambda_{21} = 1, \ \lambda_{22} = 0, \ 0 \leq \lambda_{23} \leq 1 \end{cases}$$

or

$$\begin{cases} \lambda_{11} = 0, \ 1 \leq \lambda_{12}, \ \lambda_{13} = 1 \\ 1 \leq \lambda_{21}, \ \lambda_{22} = 0, \ \lambda_{23} = 1 \end{cases}$$

for achieving the same Bayesian classifier, as well as their $E(y^*)$ and $\text{Rej}(y^*)$. The two sets of settings, however, entail different meanings to error and reject types. Hence, a confusion may be introduced when attempting to understand the behaviors of error and reject rates with respect to different sets of cost terms. For this reason, cost terms may present an intrinsic problem for defining a generic form of settings in cost-sensitive learning if a reject option is enforced.

Remark 2: It is better to apply independent parameters in the design and cost analysis of Bayesian classifiers. If no rejection, Elkan [8] proved on $N_{IP}(f, \theta_\lambda) = 1$, and suggested a single independent parameter by a cost ratio $\Lambda_r = (\lambda_{21} - \lambda_{22})/(\lambda_{12} - \lambda_{11})$. Following this principle, we suggest $\theta_\lambda = (\lambda_{11} = \lambda_{22} = 0, \lambda_{12} = 1, \lambda_{21} > 0)$ in the cost or error sensitivity analysis. A single independent cost parameter, λ_{12}, is capable of governing complete behaviors of error rate. For a reject option, we suggest $\theta_{\mathbf{T_r}}$ to be the parameters in the cost, error, or reject sensitivity analysis, which will lead to a unique interpretation to the analysis.

Remark 3: From (14) and Lemma 1, we can extend the analysis to multiple-class problems. Suppose there are m classes in classifications, its associated cost matrix, either in size of m^2 or $m(m+1)$, will have at most m independent parameters, which corresponds to the size of $\theta_{\mathbf{T_r}}$. If no distinction is made among error types, a single independent parameter ($T_{ri} = T_r, \ i = 1, \ldots, m$) will be obtained. A single independent parameter, however, does not necessarily imply an indifference among error and/or reject types (say, $T_{ri} = T_r$, $i = 1, \ldots, m-1$, $T_{rm} = 2T_r$).

Because the study of imbalanced data learning received increasing attention recently [4], [5], [7], one related theorem of Bayesian classifiers is derived in the following for elucidating their important features.

Theorem 3: In a binary classification without rejection, Bayesian classifiers with a zero–one cost function will satisfy the following rule:

$$\begin{aligned} &\text{if } p_{\min} = \min(p(t_1), p(t_2)) \to 0, \text{ and} \\ &\qquad \lambda_{11} = \lambda_{22} = 0, \lambda_{12} = \lambda_{21} = 1 \\ &\qquad\qquad \text{then } E \to E_{\max} = p_{\min} \end{aligned} \tag{16}$$

which shows that the classifiers have a tendency of reaching the UB of Bayesian error, $E_{\max}$, by misclassifying all rare-class patterns in imbalanced data learning.

Proof: Suppose that Class 2 represents a rare class. We set $p(t_1) = 1-\epsilon$ and $p(t_2) = \epsilon$ for ϵ to be an arbitrarily small positive quantity. Substituting them and $\lambda_{11} = \lambda_{22} = 0$, $\lambda_{12} = \lambda_{21} = 1$ into (6a) for a no-rejection case shows that a relation below

$$\left.\frac{p(\mathbf{x}|t_1)(1-\epsilon)}{p(\mathbf{x}|t_2)\epsilon}\right|_{\epsilon\to 0} > 1 \tag{17}$$

will always hold, so that y_1 is decided for any given pattern $\mathbf{x}$. This decision suggests that Bayesian classifiers tend to assign all patterns into the majority class in classifications when $p(t_2)$ approaches to zero. In other words, its error draws close to an UB of Bayesian error, that is, $E_{\max} = p_{\min}$. ■

C. Examples of Bayesian Classifiers on Univariate Gaussian Distributions

This section will consider abstaining Bayesian classifiers on univariate Gaussian distributions for the reason of showing closed-form solutions to the decision boundaries. When the relevant knowledge of $p(t_i)$ and $p(x|t_i)$ is given, we can depict the plots of $p(t_i|x)$ from calculation of (1) (Fig. 1). Moreover, when λ_{ij} is known, the classification regions of R_1 to R_3 in terms of x_{bj} will be fixed for Bayesian classifiers. After the regions R_1–R_3, or x_{bj}, are determined, Bayesian risk will be obtained directly. We can obtain these boundaries from the known data of δ_i when solving an equality equation on (6a) or (6b)

$$\frac{p(x = x_b|t_1)p(t_1)}{p(x = x_b|t_2)p(t_2)} = \delta_i. \tag{18}$$

The data of δ_i can be realized either from cost terms λ_{ij}, or from threshold T_{ri} [see (6)]. By substituting the exact data of $p(t_i)$ and $p(x|t_i) \sim N(\mu_i, \sigma_i)$ to Gaussian distributions, where μ_i and σ_i are the mean and standard deviation to the ith class, and the data of δ_i (for $\delta_1 = (1 - T_{r1})/T_{r1}$ from the given T_{r1}) into (18), we can obtain the closed-form solutions to the boundary points (for x_{b1} and x_{b4})

$$x_{b1,4} = \frac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2}{\sigma_1^2 - \sigma_2^2} \mp \frac{\sigma_1\sigma_2\sqrt{\alpha}}{\sigma_1^2 - \sigma_2^2}, \text{ if } \sigma_1 \neq \sigma_2 \tag{19a}$$

$$x_{b1} = \frac{\mu_1 + \mu_2}{2} + \frac{\sigma^2}{\mu_2 - \mu_1}\ln\left(\frac{p(t_1)}{p(t_2)}\frac{1}{\delta_1}\right), \text{ if } \sigma_1 = \sigma_2 = \sigma \tag{19b}$$

where α is an intermediate variable defined by

$$\alpha = (\mu_1 - \mu_2)^2 - (2\sigma_1^2 - 2\sigma_2^2)\ln\left(\frac{p(t_1)\sigma_2}{p(t_2)\sigma_1}\frac{1}{\delta_1}\right). \tag{19c}$$

Equation (19) is also effective for Bayesian classifiers in the case of no rejection. However, these four cost terms, $\lambda_{ij}(i, j = 1, 2)$, will decide the data of δ_1. The general solution to abstaining classifiers has four boundary points by substituting two thresholds T_{r1} and T_{r2}, respectively. For the conditions shown in Fig. 1(d) or (f), T_{r1} will lead to x_{b1} and x_{b4}, and T_{r2} to x_{b2} and x_{b3}, respectively. Equation (19a) shows a general form for achieving two boundary points from one data point of δ_1, and (19b) is specific for reaching a single boundary point only when the standard deviations of two classes are the same. Substituting the other data of δ_2 into (19) will yield another pair of data x_{b2} and x_{b3}, or a single one x_{b2}, in a similar form of (19).

Like the solution to boundary points, crossover point(s) can also be obtained from solving (18) or (19) by substituting $\delta_i = 1$. Three specific cases will be obtained with the crossover point(s), namely two, one, or zero crossover point(s). The case of the two crossover points appears only when $\alpha > 0$ in (19c), and two curves of $p(t_1|x)$ and $p(t_2|x)$ demonstrate the nonmonotonicity [Fig. 1(b)] with an equality relation of $p(t_1|x) = 1-p(t_2|x)$. When the associated standard deviations are equal in the two classes, i.e., $\sigma_1 = \sigma_2$, only one crossover point appears, which corresponds to the monotonous curves of $p(t_1|x)$ and $p(t_2|x)$ [Fig. 1(a)]. The case of the zero crossover point occurs when $\alpha < 0$, which corresponds to no real-value (but complex-value) solution to (19a) and situations of nonmonotonous curves of $p(t_1|x)$ and $p(t_2|x)$.

In the following, we will discuss the cases listed in Table I with respect to the number of crossover points. A term is applied to describe every case. For example, Case_k_B shows k for the kth case, and B (or M) for Bayesian (or mutual-information) classifiers.

Case_1_B (Rejection With Two Crossover Points): The necessary condition of realizing this case is

$$\frac{\lambda_{12} - \lambda_{11}}{\lambda_{21} - \lambda_{22}} < \frac{p(t_2)\sigma_1}{p(t_1)\sigma_2}e^{\frac{\mu_1 - \mu_2}{2(\sigma_1^2 - \sigma_2^2)}}. \tag{20}$$

The general rejection within this case is when $T_{r1} < 0.5$ and $1 - \max(p(t_2|x)) < T_{r2} < 0.5$, in which the reject region R_3 is divided by two ranges. When $T_{r1} < 0.5$ and $T_{r2} < 1 - \max(p(t_2|x)) < 0.5$, only one class is identified, but all other patterns are classified into a reject class. Therefore, we refer to this situation as Class 1 and reject-class classification. Table I also lists the other situations of the rejections from different settings on T_{ri}.

Case_2_B (Rejection With One Crossover Point): As shown in (19b), the general condition of this case is a simply setting $\sigma_1 = \sigma_2$. The special condition of $\alpha = 0$ in (19c) is neglected for discussions. Because the monotonicity property is enabled for the curves of $p(t_1|x)$ and $p(t_2|x)$ in this case, a single reject region is formed [Fig. 1(e)].

Case_3_B (Rejection With Zero Crossover Point): The general condition of realizing this case corresponds to a violation of the criterion on (20). In this case, one class always shows a higher value of the posterior probability distribution over the other one in the whole domain of x. From the definitions in the study of class imbalanced data set [6], [7], if $p(t_1) > p(t_2)$ in binary classifications, Class 1 will be called a majority class and Class 2 a minority class. Supposing that $p(t_1|x) > p(t_2|x)$, when $T_{r1} > 1 - \min(p(t_1|x))$, all patterns will be considered as Class 1. We call these situations a majority-taking-all classification. Because of the constraints such as $T_{r1} + T_{r2} \le 1$ and $p(t_1|x) + p(t_2|x) = 1$, one is unable to realize a minority-taking-all classification. When $T_{r1} < 1 - \min(p(t_1|x))$ and $T_{r2} < 1 - \max(p(t_2|x))$, all patterns

TABLE I

REJECTION SETTINGS OF BAYESIAN CLASSIFIERS IN UNIVARIATE GAUSSIAN DISTRIBUTIONS

$(x_{b1} < x_c < x_{b2}$ *or* $x_{b1} < x_{c1} < x_{b2} < x_{b3} < x_{c2} < x_{b4})$

Cross-over Point(s) (Reference Figure)	Rejection Thresholds	Reject region(s)	Remarks
Two (Fig. 1(d))	$T_{r1} = 0.5$, $T_{r2} = 0.5$	$\emptyset$	No Rejection
	$T_{r1} \geq 0.5$, $1 - \max(p(t_2\|x)) < T_{r2} < 0.5$	$[x_{c1}, x_{b2})$ and $(x_{b3}, x_{c2}]$	-
	$T_{r1} < 0.5$, $T_{r2} \geq 0.5$	$[x_{b1}, x_{c1})$ and $(x_{c2}, x_{b4}]$	-
	$T_{r1} < 0.5$, $1 - \max(p(t_2\|x)) < T_{r2} < 0.5$	$[x_{b1}, x_{b2})$ and $(x_{b3}, x_{b4}]$	General Rejection
	$T_{r1} < 0.5$, $T_{r2} < 1 - \max(p(t_2\|x))$	$[x_{b1}, x_{b4}]$	"*Class-1 and Reject-class*" Classification
	$T_{r1} = 0$, $T_{r2} < 1$	$(-\infty, x_{b2})$ and (x_{b3}, ∞)	"*Class-2 and Reject-class*" Classification
One (Fig. 1(e))	$T_{r1} = 0.5$, $T_{r2} = 0.5$	$\emptyset$	No Rejection
	$T_{r1} \geq 0.5$, $T_{r2} < 0.5$	$[x_c, x_{b2})$	-
	$T_{r1} < 0.5$, $T_{r2} \geq 0.5$	$[x_{b1}, x_c)$	-
	$T_{r1} < 0.5$, $T_{r2} < 0.5$	$[x_{b1}, x_{b2})$	General Rejection
Zero (Fig. 1(f))	$T_{r1} \geq 1 - \min(p(t_1\|x))$	$\emptyset$	"*Majority-taking-all*" Classification
	$T_{r1} < 1 - \min(p(t_1\|x))$ $T_{r2} < 1 - \max(p(t_2\|x))$	$[x_{b1}, x_{b4}]$	"*Majority-class and Reject-class*" Classification
	$T_{r1} < 1 - \min(p(t_1\|x))$ $T_{r2} > 1 - \max(p(t_2\|x))$	$[x_{b1}, x_{b2})$ and $(x_{b3}, x_{b4}]$	General Rejection
	$T_{r1} = 0$ $T_{r2} > 1 - \max(p(t_2\|x)) > 0.5$	$(-\infty, x_{b2})$ and (x_{b3}, ∞)	"*Minority-class and Reject-class*" Classification
Zero, one and Two (Fig.1)	$T_{r1} = T_{r2} = 0$	$(-\infty, \infty)$	Rejection to All

will be partitioned into one of two classes, that is, majority and rejection. We call these situations majority-class and reject-class classifications. The situations of minority-class and reject-class classification occur if $T_{r2} > 1 - \max(p(t_2|x)) > 0.5$ and $T_{r1} = 0$.

Case_4_B (No Rejection): In a binary classification, Chow [10] showed that, when $T_{r1} = T_{r2} \geq 0.5$, no rejection is gained to classifiers. The novel constraint of $T_{r1} + T_{r2} \leq 1$ shown in (6e) suggests that the setting should be $T_{r1} = T_{r2} = 0.5$ when the thresholds are the input data. Users need to specify an option of no rejection or rejection as an input. When no rejection is selected, the conventional scheme of cost terms from a 2 × 2 matrix will be sufficient and correct. Any usage of a 2 × 3 matrix will introduce some confusions that will be illustrated in the later section by Example 1. We cannot consider $\lambda_{13} = \lambda_{23} = 0$ as the defaults to the cost matrix in this case. Classification regions are determined by four cost terms, and then δ_i to boundary points x_{bj}. When $\delta_i = 1$ in (18), one can have a relation of $x_{bj} = x_{ci}$.

III. MUTUAL-INFORMATION CLASSIFIERS WITH A REJECT OPTION

A. Mutual-Information-Based Classifiers

Definition 5 (Mutual-Information Classifier): A mutual-information classifier is the classifier that is obtained from the maximization of mutual information over all patterns

$$y^{+} = \underset{y}{\operatorname{argmax}}\, NI(T = t; Y = y) \tag{21}$$

where T and Y are the target variable and decision output variable, t and y are their values, respectively. For simplicity, we denote $NI(T = t; Y = y) = NI(T; Y)$ as the NI in the following form [28]:

$$\mathrm{NI}(T; Y) = \frac{I(T; Y)}{H(T)} \tag{22a}$$

where $H(T)$ is the entropy based on the Shannon definition [21] to the target variable

$$H(T) = -\sum_{i=1}^{m} p(t_i) log_2 p(t_i) \tag{22b}$$

and $I(T; Y)$ is the mutual information between two variables of T and Y [21]

$$I(T; Y) = \sum_{i=1}^{m} \sum_{j=1}^{m+1} p(t_i, y_j) \log_2 \frac{p(t_i, y_j)}{p(t_i)p(y_j)} \tag{22c}$$

where m is the total number of classes in T. In the case without rejection, (22c) will be realized by a summation of j over 1 to m, instead of to $m + 1$. $p(t, y)$ is the joint distribution between the two variables, and $p(t)$ and $p(y)$ are the marginal distributions, which can be derived from [21]

$$p(t) = \sum_{y} p(t, y), \quad p(y) = \sum_{t} p(t, y). \tag{23}$$

Considering binary classifications with a reject option, one will have the following formula to the joint distribution $p(t, y)$:

$$p(t, y) = \begin{bmatrix} \int_{R_1} p(t_1)p(\mathbf{x}|t_1)d\mathbf{x} & \int_{R_2} p(t_1)p(\mathbf{x}|t_1)d\mathbf{x} & \int_{R_3} p(t_1)p(\mathbf{x}|t_1)d\mathbf{x} \\ \int_{R_1} p(t_2)p(\mathbf{x}|t_2)d\mathbf{x} & \int_{R_2} p(t_2)p(\mathbf{x}|t_2)d\mathbf{x} & \int_{R_3} p(t_2)p(\mathbf{x}|t_2)d\mathbf{x} \end{bmatrix}. \tag{24}$$

The marginal distribution of $p(t)$ is in fact the given information of prior knowledge about the classes

$$p(t) = (p(t_1), p(t_2))^T \tag{25}$$

where the superscript T represents a transpose, and the marginal distribution of $p(y)$ is as follows:

$$p(y) = (p(y_1), p(y_2), p(y_3)) = \left(\int\limits_{R_1} Q d\mathbf{x}, \int\limits_{R_2} Q d\mathbf{x}, \int\limits_{R_3} Q d\mathbf{x} \right)$$
$$Q = p(t_1)p(\mathbf{x}|t_1) + p(t_2)p(\mathbf{x}|t_2). \tag{26}$$

Substituting (24) and (25) into (21), we can describe NI in terms of $p(t_i)$ and $p(\mathbf{x}|t_i)$. A normalization scheme is applied so that a relative comparison can be made among classifiers. When the prior knowledge of $p(t_i)$ is given, the entropy $H(T)$ in (22b) will be unchanged during classifier learning. This is why we use this term to normalize the mutual information in (22a).

Remark 4: When processing real-world data, both mutual-information and Bayesian classifiers can adopt a confusion matrix for searching their final solutions, but with different objective functions. Mathematically, (21) expresses that y^+ is an optimal classifier in terms of the maximal mutual information, or relative entropy, between the target variable T and decision output variable Y. The physical interpretation of relative entropy is a measurer of probability similarity between the two variables. In a multiple-class classification with m classes, if an $m \times m$ confusion matrix is formed, it implies a no rejection for both types of classifiers (i.e., only R_1–R_m are determined). If an $m \times (m+1)$ confusion matrix is given, it will conduct a classification with a reject option (i.e., R_1–R_{m+1} are formed). Users make a selection of this option through specifying the size of the confusion matrix.

Definition 6 (Augmented Confusion Matrix [28]): An augmented confusion matrix will include one column for a rejected class, which is added on a conventional confusion matrix

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} & c_{1(m+1)} \\ c_{21} & c_{22} & \cdots & c_{2m} & c_{2(m+1)} \\ & & \cdots & & \\ c_{m1} & c_{m2} & \cdots & c_{mm} & c_{m(m+1)} \end{bmatrix} \tag{27}$$

where c_{ij} is the number of the ith class that is classified as the jth class. The row data correspond to the exact classes, and the column data correspond to the prediction classes. The last column represents a reject class. The relations and constraints of an augmented confusion matrix are

$$C_i = \sum_{j=1}^{m+1} c_{ij}, \ C_i > 0, \ c_{ij} \geq 0, \quad i = 1, 2, \ldots, m \tag{28}$$

where C_i is the total number of the ith class. The data of C_i are known in classification problems.

In this paper, supposing that the input data of classifications are exactly known about the prior probability $p(t_i)$ and the conditional probability density function $p(\mathbf{x}|t_i)$, one is able to derive the joint distribution in association with the confusion matrix

$$p_{ij} = p(t_i, y_j) = \int\limits_{R_j} p(t_i)p(\mathbf{x}|t_i)d\mathbf{x} \approx \frac{c_{ij}}{N} = p_e(t_i, y_j)$$
$$i = 1, 2, \ldots, m, \ j = 1, 2, \ldots, m+1. \tag{29}$$

where R_j is the region in which every pattern $\mathbf{x}$ is identified as the jth class, and $p_e(t_i, y_j)$ is the empirical probability density in applications where only a confusion matrix is given. In those applications, the total number of patterns N is generally known.

Equation (28) describes the approximation relations between the joint distribution and confusion matrix. If the knowledge about $p(t_i)$ and $p(\mathbf{x}|t_i)$ are exactly known, we can design a mutual-information classifier directly. If no initial information is known about $p(t_i)$ and $p(\mathbf{x}|t_i)$, the empirical probability density of joint distribution, $p_e(t_i, y_j)$, can be estimated from the confusion matrix [28]. This treatment, based on the frequency principle of a confusion matrix, is not mathematically rigorous, but will offer a simple approach for classifiers to apply the entropy principle in wider applications.

B. Examples of Mutual-Information Classifiers on Univariate Gaussian Distributions

Mutual-information classifiers, like Bayesian classifiers, also provide a general formulation to classifications. They are able to process classifications with or without rejection. This section will aim at deriving novel formulas necessary in the design and analysis of mutual-information classifiers under assumptions of Gaussian distributions. The input data are the same as those of Bayesian classifiers shown in Section II, except that cost terms of λ_{ij} are not given as the input, but will be displayed as the output of the classifiers. In other words, mutual-information classifiers will automatically calculate the two thresholds that can lead to the cost terms through (7). However, because of a redundancy among six cost terms, we will fail to obtain the unique solution of the cost terms, which is demonstrated in Example 1 of Section IV.

Generally, one is unable to derive a closed-form solution to mutual-information classifiers. One of the obstacles is the nonlinear complexity of solving error functions. Therefore, this paper only provides semianalytical solutions to mutual-information classifiers. When substituting $p(t_i)$ and $p(x|t_i)$ into (21), we will encounter the process of solving an inverse problem on the following function:

$$\max_{y \in Y} NI(T; Y) = \max f(x, \theta = (p(t_i), p(x|t_i), x_{bj})) \tag{30}$$

for searching the boundary points x_{bj} from error functions. Only numerical solutions can be obtained to x_{bj}, except for a special case. In the following, some specific cases in Fig. 1 will be discussed on mutual-information classifiers in related to the number of crossover points.

Case_1_M (Rejection With Two Crossover Points): This is a general case of mutual-information classifiers in which four boundary points, x_{bj}, are formed [Fig. 1(d)]. When the four points obtained numerically from solving (30), the

$$p(t,y) = \begin{bmatrix} \frac{p(t_1)}{2}[1-erf(X_{11})] & \frac{p(t_1)}{2}[1-erf(X_{12})] & \frac{p(t_1)}{2}[erf(X_{11})+erf(X_{12})] \\ \frac{p(t_2)}{2}[1-erf(X_{21})] & \frac{p(t_2)}{2}[1-erf(X_{22})] & \frac{p(t_2)}{2}[erf(X_{21})+erf(X_{22})] \end{bmatrix} \tag{31a}$$

classification regions R_1–R_3 will be made. With the condition of $x_{b1} < x_{b2} < x_{b3} < x_{b4}$, the closed-form solution of $p(t, y)$ can be given in a form of (31a), as shown at the top of this page, where erf(·) is an error function, and

$$X_{ij} = \frac{\mu_i - x_{bj}}{\sqrt{2}\sigma_i}, \quad i = 1, 2, \ j = 1, 2. \tag{31b}$$

With (31a), we can get the error rate and reject rate from

$$\begin{aligned} E &= E_1 + E_2 = p(t_i = 1, y_j = 2) + p(t_i = 2, y_j = 1) \\ &= \frac{p(t_1)}{2}[1 - \mathrm{erf}(X_{12})] + \frac{p(t_2)}{2}[1 - \mathrm{erf}(X_{21})] \end{aligned} \tag{32a}$$

$$\begin{aligned} \mathrm{Rej} &= \mathrm{Rej}_1 + \mathrm{Rej}_2 \\ &= p(t_i = 1, y_j = 3) + p(t_i = 2, y_j = 3) \\ &= \frac{p(t_1)}{2}[\mathrm{erf}(X_{11}) + \mathrm{erf}(X_{12})] \\ &\quad + \frac{p(t_2)}{2}[\mathrm{erf}(X_{21}) + \mathrm{erf}(X_{22})]. \end{aligned} \tag{32b}$$

The rejection thresholds are also derived from the given data of x_{bj}

$$\begin{aligned} T_{r1} &= 1 - p(t_1|x = x_{b1}) \\ &= 1 - \frac{p(t_1)\sigma_2 e^{\frac{-(x_{b1}-\mu_1)^2}{2\sigma_1^2}}}{p(t_1)\sigma_2 e^{\frac{-(x_{b1}-\mu_1)^2}{2\sigma_1^2}} + p(t_2)\sigma_1 e^{\frac{-(x_{b1}-\mu_2)^2}{2\sigma_2^2}}} \end{aligned} \tag{33a}$$

$$\begin{aligned} T_{r2} &= 1 - p(t_2|x = x_{b2}) \\ &= 1 - \frac{p(t_2)\sigma_1 e^{\frac{-(x_{b2}-\mu_2)^2}{2\sigma_2^2}}}{p(t_1)\sigma_2 e^{\frac{-(x_{b2}-\mu_1)^2}{2\sigma_1^2}} + p(t_2)\sigma_1 e^{\frac{-(x_{b2}-\mu_2)^2}{2\sigma_2^2}}}. \end{aligned} \tag{33b}$$

With the condition of $x_{b1} < x_{b2} < x_{b3} < x_{b4}$ shown in Fig. 1(d), substituting either x_{b1} or x_{b4} into (33) will give the same value on T_{r1}, and a similar one to x_{b2} or x_{b3} on T_{r2}. The results of T_{r1} and T_{r2} show that mutual-information classifiers will automatically search the rejection thresholds for balancing the error and reject rates from the given data of classes. When T_{r1} and T_{r2} are known, a mutual-information classifier can have its equivalent Bayesian classifiers through the relations of (7). This specific feature will be discussed in Section IV.

Case_2_M (Rejection With One Crossover Point): Mutual-information classifiers are able to calculate two boundary points x_{b1} and x_{b2}, or a single reject region [Fig. 1(e)], from the given data.

Case_3_M (Rejection With Zero Crossover Point): This case shows a similar result to that in the case with two crossover points.

Case_4_M (No Rejection): Mutual-information classifiers will calculate boundary points accordingly [Fig. 1(a)–(c)]. A very special case appears in which one is able to obtain a closed-form solution to mutual-information classifiers. This case corresponds to the conditions of $p(t_1) = p(t_2)$ and $\sigma_1 = \sigma_2$, for no rejection. The solution shows a single boundary point x_b, coincident to the crossover point x_c, for partitioning the classification regions

$$\begin{aligned} x_b = x_c &= \frac{\mu_1 + \mu_2}{2} \\ \text{if } \mu_1 < \mu_2 \text{ then } R_1 &= (-\infty, x_b),\ R_2 = [x_b, \infty),\ R_3 = \emptyset. \end{aligned} \tag{34}$$

This result is the same of Bayesian classifiers, which leads to the same error rates between the two types of classifiers. The special case describes a relation of $y^+ = y^*$ only under strictly limited conditions.

IV. Comparisons Between Bayesian Classifiers and Mutual-Information Classifiers

A. General Comparisons

Mutual-information classifiers provide users a new perspective in processing classification problems, hence enlarge the toolbox in their applications. For discovering new features in this approach, this section will discuss general aspects of mutual-information and Bayesian classifiers simultaneously for a systematic comparison. The main objective of the comparative study is to reveal their corresponding advantages and disadvantages. Meanwhile, their associated issues, or new challenges, are also presented from the personal viewpoint of the author.

First, both types of classifiers share the same assumptions of requiring the exact knowledge about class distributions and specifying the status of the reject option (Table II). The exact knowledge feature imposes the most weakness on the two approaches in applications. In other words, the approaches are considered more theoretically meaningful, rather than directly useful in solving real-world problems. When the exact knowledge is not available, the existing estimation approaches to class distributions [2], [33] in Bayesian classifiers will be feasible for implementing mutual-information classifiers. The learning targets of Bayesian classifiers involve evaluations of risks or errors, which are mostly compatible with classification goals in real-life applications. However, the concept of mutual information, or entropy-based criteria, is not a common concern or requirement from most of the classifier designers and users [28].

Second, Bayesian classifiers will ask (or implicitly apply) cost terms for their designs. This requirement provides both

TABLE II

DATA INFORMATION OF BAYESIAN AND MUTUAL-INFORMATION CLASSIFIERS IN BINARY CLASSIFICATIONS

Classifier	Required Input		Learning	Output
Type	On Data	On Rejection	Target	Data
Bayesian	$p(t_1)$, $p(t_2)$ $p(\mathbf{x} \mid t_1)$, $p(\mathbf{x} \mid t_2)$ λ_{11}, λ_{12}, λ_{13} λ_{21}, λ_{22}, λ_{23}	No or Yes	$\min Risk(y)$ or $\min E(y)$	E_1, E_2, Rej_1, Rej_2, $Risk$, R_1, R_2, R_3, T_{r1}, T_{r2}
Mutual- Information	$p(t_1)$, $p(t_2)$ $p(\mathbf{x} \mid t_1)$, $p(\mathbf{x} \mid t_2)$	No or Yes	$\max NI(T;Y)$	E_1, E_2, Rej_1, Rej_2, NI, R_1, R_2, R_3, T_{r1}, T_{r2} ($\{\lambda_{21}/\lambda_{12}\}$, or $\{\lambda_{21}, \lambda_{13}, \lambda_{23}\}$)

advantages and disadvantages depending on the applications. The main advantage is its flexibility in offering objective or subjective designs of classifiers. When the exact knowledge is available and reliable, inputting such data will be very simple and meaningful for realizing objective designs. Simultaneously, subjective designs will always be possible. The main disadvantage may occur in objective designs if one has incomplete information about cost terms. Generally, cost terms are more liable to subjectivity than prior probabilities. In this case, avoiding the introduction of subjectivity is not an easy task for Bayesian classifiers. Mutual-information classifiers, without requiring cost terms, will fall into an objective approach. They carry an intrinsic feature of letting the data speak for itself, which exhibits a significant difference from a subjective version of Bayesian classifiers. The current definition of mutual-information classifiers, however, needs to be extended for carrying the flexibility of subjective designs, which is technically feasible by introducing free parameters, such as fuzzy entropy [34].

Third, one of the problems from the current learning targets of Bayesian classifiers is their failure to obtain the optimal rejection threshold in classifications. Although Chow [10] and Ha [18] suggested formulas, respectively, in the forms of

$$\min \mathrm{Risk}(T_r) = E(T_r) + T_r \mathrm{Rej}(T_r) \tag{35a}$$

or

$$\min \frac{E(T_r)}{\mathrm{Rej}(T_r)} \tag{35b}$$

a minimization from both formulas will lead to a solution of $T_r = 0$ for Risk $= 0$, which implies a rejection of all patterns. Therefore, we are expecting to establish a learning target of determining optimal rejection thresholds of Bayesian classifiers. Information-based classifiers seem to be unique for achieving the optimal rejection thresholds as the classifiers' outcomes. The remaining issue is to study them in a systematic way.

Fourth, Bayesian classifiers generally fail to handle the class-imbalanced data properly if no cost terms are specified in classifications, as described in Theorem 3. When one class approximates a smaller (or zero) population and no distinction is made among error types, Bayesian classifiers have a tendency to put all patterns of the smaller class into error, and its NI will be approximately zero, which represents that no information is obtained from classifiers [19]. Mutual-information classifiers display unique advantages in these situations, including cases of abstaining classifications. They provide a solution of balancing error and reject types without using cost terms. The challenge lies in their theoretical derivation of response behaviors, such as UB and lower bound (LB) of $E_i/p(t_i)$ for mutual-information classifiers.

Fifth, mutual-information classifiers will add extra computational complexities and costs over Bayesian classifiers. Both types of classifiers require computations of posterior probability. When these data are obtained, Bayesian classifiers will produce decision results directly. Mutual-information classifiers will, however, need further procedures, such as to form a confusion matrix (or a joint distribution), evaluate NI in (22), and search boundary points from a nonconvex space NI in (30). These procedures will introduce significantly analytical and computational difficulties to mutual-information classifiers, particularly in multiple-class problems with high dimensions.

Note that the above-mentioned discussions provide a preliminary answer to the question posed in the title of this paper. In another connection, Appendix B presents the bounds between conditional entropy and Bayesian error in binary classifications. Further investigations are expected to search other differences under various assumptions or backgrounds, such as distributions of mixture models, multiple-class classifications in high-dimension variables, rejection to a subset of classes [18], and experimental studies from real-world data sets.

B. Comparisons on Univariate Gaussian Distributions

Gaussian distributions are not only important in theoretical sense but also, to a large extent, appropriate for providing critical guidelines in real applications. In classification problems, many important findings can be revealed from a study on Gaussian distributions.

The following numerical examples are specifically designed for demonstrating the intrinsic differences between Bayesian and mutual-information classifiers on Gaussian distributions.

TABLE III
RESULTS OF EXAMPLE 1 ON UNIVARIATE GAUSSIAN DISTRIBUTIONS

Reject	Classifier	E_1	E	Rej_1	Rej	T_{r1}	x_{b1}, x_{b2}	NI
Option	Type	E_2		Rej_2		T_{r2}	x_{b3}, x_{b4}	
No Rejection	Bayesian	0.170	**0.227**	0	0	-	-0.238, 3.571	0.245
		0.057		0		-	-, -	
	Mutual-	0.215	0.239	0	0	-	-0.674, 4.007	**0.260**
	Information	0.024		0		-	-, -	
Rejection	Bayesian	0.131	**0.155**	0.083	0.167	0.333	-0.673, 0.162	0.285
		0.024		0.084		0.375	3.171, 4.006	
	Mutual-	0.154	0.160	0.118	0.186	0.141	-1.24, -0.0762	**0.297**
	Information	0.006		0.068		0.445	3.409, 4.571	

Example 1 (Two Crossover Points): The data for no rejection are given in the following:

No rejection:

$$\mu_1 = -1,\ \sigma_1 = 2,\ p(t_1) = 0.5,\ \lambda_{11} = 0,\ \lambda_{12} = 1$$

$$\mu_2 = 1,\quad \sigma_2 = 1,\ p(t_2) = 0.5,\ \lambda_{21} = 1,\ \lambda_{22} = 0.$$

The cost terms are used in Bayesian classifiers, but not in mutual-information classifiers. Table III lists the results of both classifiers. We can obtain the same results when inputing $\lambda_{13} = 1 - \lambda_{23}$ in Bayesian classifiers. This is why a 2×2 matrix has to be used in the case of no rejection. Two crossover points are formed in this example [Fig. 1(a)]. If no rejection is selected, both classifiers will have two boundary points. Bayesian classifiers will partition the classification regions by having $x_{b1} = x_{c1} = -0.238$ and $x_{b2} = x_{c2} = 3.57$. Mutual-information classifiers widen the region R_2 by $x_{b1} = -0.674$ and $x_{b2} = 4.007$ so that the error of Class 2 is much reduced. If considering zero costs for correct classifications and using (18), we can calculate a cost ratio below as an independent parameter to Bayesian classifiers in the case of no rejection

$$\Lambda_r = \frac{\lambda_{21}}{\lambda_{12}} = \frac{p(x = x_b|t_1)p(t_1)}{p(x = x_b|t_2)p(t_2)} \tag{36}$$

which is used to establish an equivalence between mutual-information and Bayesian classifiers. Substituting the boundary points of mutual-information classifier at $x_{b1} = -0.674$ and $x_{b2} = 4.007$ into $p(x|t_i)$ and (36), respectively, we receive a unique cost ratio value, $\Lambda_r = 2.002$. Hence, this mutual-information classifier has its unique equivalence to a specific Bayesian classifier, which is exerted by the following conditions to the cost terms:

$$\lambda_{11} = 0,\ \lambda_{12} = 1.0,\ \lambda_{21} = 2.002,\ \lambda_{22} = 0.$$

When two classes are well balanced, that is, $p(t_1) = p(t_2)$, both types of classifiers will produce larger errors in association with the larger variance class. Mutual-information classifiers, however, always add more cost weight ($\lambda_{21} = 2.002$) on the misclassification from a smaller variance class. In other words, mutual-information classifiers prefer to generate a smaller error on a smaller variance class in comparison with Bayesian classifiers when using a zero–one cost function (Table III). This performance behavior seems closer to our intuitions in binary classifications under the condition of a balanced class data set. When two classes are significantly different from their associated variances, a smaller variance class generally represents an interested signal embedded within noise, which often has a larger variance. The common practices in such classification scenarios require a larger cost weight on the misclassification from a smaller variance class, and vice versa from a larger variance class.

If a reject option is enforced by the following data.

Rejection:

$$\mu_1 = -1,\ \sigma_1 = 2,\ p(t_1) = 0.5$$

$$\mu_2 = 1,\quad \sigma_2 = 1,\ p(t_2) = 0.5$$

$$\lambda_{11} = 0,\ \lambda_{12} = 1.2,\ \lambda_{13} = 0.2$$

$$\lambda_{21} = 1,\ \lambda_{22} = 0,\quad \lambda_{23} = 0.6.$$

Four boundary points are required to determine classification regions, as shown in Fig. 1(d). For the given cost terms, a Bayesian classifier shows a lower error rate and a lower reject rate in comparison with its counterpart. While the rejects are almost equal between two classes, the errors are significantly different. One is able to adjust the errors and rejects by changing cost terms. From a mutual-information classifier, a balance is automatically made among error and reject types. The results, shown in Table III, are considered for carrying the feature of objectivity in evaluations as no cost terms are specified subjectively. Note that a reject option enables both types of classifiers to reach higher values on their NIs than those without rejection. Because no one-to-one relations exist among the thresholds and the cost terms in a rejection case, one will fail to acquire a unique set of the equivalent cost terms between the Bayesian and the mutual-information classifiers. For example, two sets of cost terms in the following will produce the same Bayesian classifiers based on the given solutions of the mutual-information classifier

$$\begin{cases} \lambda_{11} = 0,\ \lambda_{12} = 1,\ \lambda_{13} = 0.0376 \\ \lambda_{21} = 1,\ \lambda_{22} = 0,\ \lambda_{23} = 0.772 \end{cases}$$

TABLE IV

RESULTS OF EXAMPLE 2 ON UNIVARIATE GAUSSIAN DISTRIBUTIONS

Classifier Type	$p(t_1)/p(t_2)$ $[p(t_1), p(t_2)]$	1 $[0.5, 0.5]$	2 $[2/3, 1/3]$	4 $[0.8, 0.2]$	9 $[0.9, 0.1]$	99 $[0.99, 0.01]$	999 $[0.999, 0.001]$	9999 $[0.9999, 0.0001]$
Bayesian	E_1	0.0793	0.0594	0.0362	0.0161	0.483e-3	0.422e-5	0.000
	E_2	0.0793	0.0856	0.0759	0.0539	0.903e-2	0.993e-3	0.1e-3
	$E_2/p(t_2)$	0.159	0.257	0.379	0.539	0.903	0.993	1.000
	$x_b(=x_c)$	0.0	0.347	0.693	1.10	2.30	3.45	4.61
	$H(T\|Y)$	0.631	0.591	0.491	0.349	0.0756	0.0113	0.00147
	NI	0.369	0.356	0.320	0.256	0.0644	0.00524	0.124e-3
Mutual-Information	E_1	0.0793	0.0867	0.0852	0.0772	0.0585	0.0551	0.0547
	E_2	0.0793	0.0637	0.0451	0.0264	0.331e-2	0.343e-3	0.345e-4
	$E_2/p(t_2)$	0.159	0.191	0.225	0.264	0.331	0.343	0.345
	x_b	0.0	0.126	0.246	0.367	0.562	0.597	0.601
	$H(T\|Y)$	0.631	0.586	0.472	0.320	0.0629	0.00957	0.00129
	NI	0.369	0.362	0.346	0.317	0.222	0.161	0.125

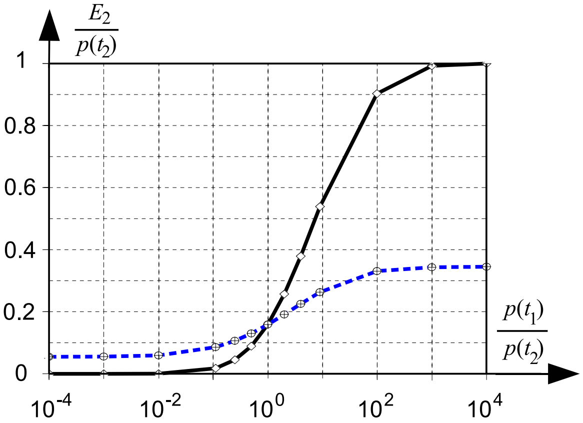


Fig. 2. Curves of $E_2/p(t_2)$ versus $p(t_1)/p(t_2)$ for Example 2. Solid curve: Bayesian classifier. Dashed curve: Mutual-information classifier.

or

$$\begin{cases} \lambda_{11} = 0, \ \lambda_{12} = 2.247, \ \lambda_{13} = 1 \\ \lambda_{21} = 7.069, \ \lambda_{22} = 0, \ \lambda_{23} = 1. \end{cases}$$

The meanings for two sets of cost terms are different. The first set shows the same costs for errors, but the second one suggests the same costs for rejects. The above-mentioned results imply an intrinsic problem of nonconsistency for interpreting cost terms. One needs to be cautious about this problem when setting cost terms to Bayesian classifiers. This phenomenon occurs only in the case that a reject option is considered, but does not in the case without rejection. If the knowledge about thresholds exists, abstaining classifiers are better to apply T_{ri} directly as the input data, instead of employing cost terms. If no information is given about the thresholds or cost terms, mutual-information classifiers are able to provide an objective, or initial, reference of T_{ri} for Bayesian classifiers in cost-sensitive learning.

Example 2 (One Crossover Point): The given inputs in this example are as follows.

No rejection:
$\mu_1 = -1, \ \sigma_1 = 1, \ \lambda_{11} = 0, \ \lambda_{12} = 1$
$\mu_2 = 1, \ \sigma_2 = 1, \ \lambda_{21} = 1, \ \lambda_{22} = 0$
$p(t_1) = 0.5, 2/3, 0.8, 0.9, 0.99, 0.999, 0.9999$
$p(t_2) = 0.5, 1/3, 0.2, 0.1, 0.01, 0.001, 0.0001.$

Specific attention is paid to the class imbalanced data. When Class 2 alters from balanced, minority to rare status in the whole data, we need to find out what behaviors both types of classifiers will display. For this purpose, a natural scheme with zero–one cost terms is set to Bayesian classifiers. Numerical investigations are conducted in this example. Table IV lists the results of classifiers on the given data. If following the conventional term FNR for false negative rate in binary classifications, which is defined as

$$\text{FNR} = \frac{E_2}{p(t_2)} \tag{37}$$

one can examine behaviors of FNR with respect to the ratio $p(t_1)/p(t_2)$. Sometimes, FNR is also called a miss rate [2]. Two types of classifiers show the same results when two classes are exactly balanced, that is, $p(t_1)/p(t_2) = 1$. A single boundary point [Fig. 1(b)] separates two classes at the exact crossover point ($x_b = x_c = 0$) according to (36). When one class, say $p(t_2)$ for Class 2, becomes smaller, the boundary point of Bayesian classifier moves toward to the mean point ($\mu_2 = 1$) of Class 2 [2, p. 39], and passes it finally. For keeping the smallest error, a Bayesian classifier will sacrifice the minority class. The results in Table IV confirm Theorem 3 numerically on the Bayesian classifiers. Fig. 2 shows such behavior from the plot of $E_2/p(t_2)$ versus $p(t_1)/p(t_2)$. Note that the plots in the range from 10^{-4} to 10^0 on the $p(t_1)/p(t_2)$ axis are also shown based on the data in Table IV. For example, at the data point of $p(t_1)/p(t_2) = 1/2$, one can get $E_2/p(t_2) = 0.0594/(2/3)$, where 0.0594 is taken from E_1 for the data at $p(t_1)/p(t_2) = 2$. One can observe that the complete set of Class 2 could be misclassified when it becomes extremely rare. This finding explains another reason for the question: Why do classifiers perform worse on the minority class? in [39].

Mutual-information classifiers exhibit different behavior in the given data set. The first important feature is that the boundary point will shift toward the mean point ($\mu_2 = 1$) of Class 2 but will never go over it. The second feature informs that the response of $E_2/p(t_2)$ approaches asymptotically to a stable value, about 0.345 in this example, for a large ratio of $p(t_1)/p(t_2)$. This feature shows that mutual-information

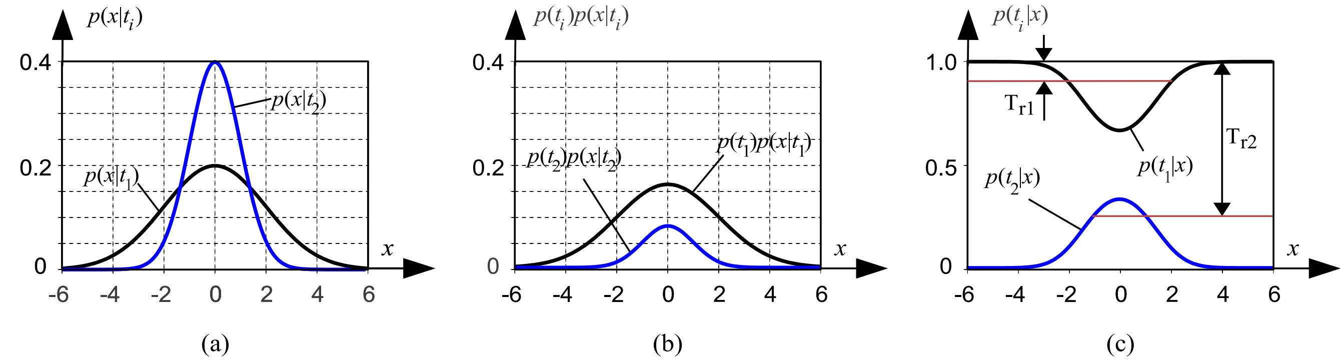


Fig. 3. Plots of Example 3 where (b) and (c) describe a signal (blue curve) embedded by wider band noise (black curve). T_{r1} and T_{r2} are the rejection thresholds of Classes 1 and 2, respectively.

TABLE V
RESULTS OF EXAMPLE 3 ON UNIVARIATE GAUSSIAN DISTRIBUTIONS

Reject Option	Classifier Type	E_1 E_2	E	Rej_1 Rej_2	Rej	T_{r1} T_{r2}	x_{b1}, x_{b2} x_{b3}, x_{b4}	NI
No Rejection	Bayesian	0.0 0.2	0.2	0 0	0	- -	-, - -, -	0.0
	Mutual-Information	0.499 0.0153	0.514	0 0	0	- -	-1.77, 1.77 -, -	0.0803
Rejection	Mutual-Information	0.316 0.00819	0.324	0.239 0.0520	0.291	0.0945 0.749	-2.04, -1.03 1.03, 2.04	**0.0926**

classifiers will automatically increase the cost weight to a class who becomes rare. A significant fraction of the rare class, 65.5%, is identified correctly. Moreover, the curve of $E_2/p(t_2)$ also demonstrates a lower, yet nonzero, bound on error rate (about 0.054) when $p(t_1)/p(t_2)$ approaches to zero.

Example 2 demonstrates different interpretations behind learning targets. If a decision rule of less costs more is applied in classifications, the plots from Fig. 2 will advocate for mutual-information classifiers, rather than for Bayesian classifiers. From a theoretical viewpoint, we still, however, need to establish an analytical relation of the stable points for mutual-information classifiers.

Example 3 (Zero Crossover Points): The given data of two classes are as follows:

$$\mu_1 = 0,\ \ \sigma_1 = 2,\ \ p(t_1) = 0.8$$
$$\mu_2 = 0,\ \ \sigma_2 = 1,\ \ p(t_2) = 0.2.$$

Although no data are specified to the cost terms, it generally implies a zero–one lost function for them [2]. From (19c), one can see a case of zero crossover point occurs in this example [Fig. 3(c)]. In the zero–one setting to cost terms, the Bayesian classifier will produce a specific classification result of majority-taking-all, that is, for all patterns identified as Class 1. The Bayesian error gives to Class 2 only, and the relation of $NI = 0$ shows that no information is obtained from the classifier. One can imagine that the given example may describe a classification problem where a target class, with Gaussian distribution, is fully corrupted with wider band Gaussian noise in a frequency domain [Fig. 3(b)]. The plots of $p(t_i)p(x|t_i)$ show the overwhelming distribution of Class 1 over that of Class 2. Hence, the plots on the posterior probability $p(t_i|x)$ show that Class 2 has no chance to be considered in the complete domain of x [Fig. 3(c)].

Table V lists the results for both types of classifiers. The Bayesian approach fails to achieve the meaningful results on the given data. When missing input data of λ_{13} and λ_{23}, one cannot carry out the Bayesian approach in abstaining classifications. On the contrary, without specifying any cost term, mutual-information classifiers are able to detect the target class with a reasonable degree of accuracy. When no rejection is selected, less than 2 percentage error ($E2 = 0.0153$) happens to the target class. Although the total error ($E = 0.514$) is much higher than its Bayesian counterpart ($E = 0.200, \mathrm{FNR} = 0.0$), the result about a lower miss rate (FNR $= 0.0765$) to the target is really meaningful in applications. If a reject option is engaged, the miss rate is further reduced to FNR $= 0.0410$, but adds a reject rate of Rej $= 0.291$ over total possible patterns. This example confirms again the unique feature of mutual-information classifiers. The results of the cost ratio ($\Lambda_r = 6.475$) in the case of no rejection from mutual-information classifiers can serve a useful reference for a cost-sensitive learning when missing information about costs.

V. CONCLUSION

This paper explored differences between Bayesian and mutual-information classifiers. With Chow's pioneering work [9], [10], the author revisited Bayesian classifiers on two general scenarios for the reason of their increasing popularity in classifications. The first was on a zero–one cost function to classifications without rejection. The second was on the distinctions among error and reject types in abstaining classifications. In addition, this paper focused on the analytical study of mutual-information classifiers in comparison with Bayesian classifiers, which showed a basis for novel design or analysis of classifiers based on the entropy principle. The general decision rules were derived for both Bayesian and mutual-information classifiers based on the given assumptions. Two specific theorems were derived for revealing the intrinsic problems of Bayesian classifiers in applications under the two scenarios. One theorem described that Bayesian classifiers have a tendency of overlooking the misclassification error which is associated with a minority class. This tendency will degenerate a binary classification into a single-class problem with the meaningless solutions. The other theorem discovered the parameter redundancy of cost terms in abstaining classifications. This weakness is not only on reaching an inconsistent interpretation to cost terms. The pivotal difficulty will be on holding the objectivity of cost terms. In real applications, information about cost terms is rarely available. This is particularly true for reject types. In comparison, mutual-information classifiers do not suffer such difficulties. Their advantages without requiring cost terms will enable the current decision systems to process abstaining classifications in an objective means. Several numerical examples in this paper supported the unique benefits of using mutual-information classifiers in special cases.

The comparative study in this paper was not meant to replace Bayesian classifiers by mutual-information classifiers. Both types of classifiers can form complementary solutions, such as cost-free learning [40], [41] in difference from cost-sensitive learning [42] for real-world problems. This paper was intended to highlight their differences. More detailed discussions to the differences between the two types of classifiers were given in Section IV. As a final conclusion, a simple answer to the question title is summarized below.

Bayesian and mutual-information classifiers are different essentially from their learning targets applied. From the application viewpoints, Bayesian classifiers are more suitable to the cases when cost terms are exactly known for tradeoff of error and reject types. Mutual-information classifiers are capable of objectively balancing error and reject types automatically without employing cost terms, even in the cases of extremely class-imbalanced data sets, which may describe a theoretical interpretation why humans are more concerned about the accuracy of rare classes in classifications.

Appendix I
Proof of Theorem 1

Proof: The decision rule of Bayesian classifiers in the no-rejection case is well known in [2]. Then, only the rule of the rejection case is studied in this present proof. Considering (6a) first from (5a), a pattern $\mathbf{x}$ is decided by a Bayesian classifier to be y_1 if $\text{risk}(y_1|\mathbf{x}) < \text{risk}(y_2|\mathbf{x})$ and $\text{risk}(y_1|\mathbf{x}) < \text{risk}(y_3|\mathbf{x})$. Substituting (1) and (2) into these inequality equations will result to

$$\text{Decide } y_1 \text{ if } \frac{p(\mathbf{x}|t_1)p(t_1)}{p(\mathbf{x}|t_2)p(t_2)} > \frac{\lambda_{21}-\lambda_{22}}{\lambda_{12}-\lambda_{11}}$$
$$\text{and } \frac{p(\mathbf{x}|t_1)p(t_1)}{p(\mathbf{x}|t_2)p(t_2)} > \frac{\lambda_{21}-\lambda_{23}}{\lambda_{13}-\lambda_{11}}. \tag{A1}$$

Similarly, one can obtain

$$\text{Decide } y_2 \text{ if } \frac{p(\mathbf{x}|t_1)p(t_1)}{p(\mathbf{x}|t_2)p(t_2)} \leq \frac{\lambda_{21}-\lambda_{22}}{\lambda_{12}-\lambda_{11}}$$
$$\text{and } \frac{p(\mathbf{x}|t_1)p(t_1)}{p(\mathbf{x}|t_2)p(t_2)} \leq \frac{\lambda_{23}-\lambda_{22}}{\lambda_{12}-\lambda_{13}} \tag{A2}$$

and (6c), respectively. (A1) describes that a single UB within two boundaries will control a pattern $\mathbf{x}$ to be y_1. Similarly, (A2) describes an LB for a pattern $\mathbf{x}$ to be y_2. From the constraints in (3), one cannot determine which boundaries will be UB or LB. However, one can determine them from the following two hints in classifications.

1) (6c) describes a single lower boundary and a single upper boundary for a pattern $\mathbf{x}$ to be y_3.
2) The UB in (A1) and the LB in (A2) should be coincident with one of the boundaries in (6c), respectively, so that the classification regions from R_1 to R_3 will cover a complete domain of the pattern $\mathbf{x}$ [Fig. 1(d)–(f)].

The above-mentioned hints suggest the novel constraints of λ_{ij}, as shown in (6d). Any violation of the constraints will introduce a new classification region R_4, which is not correct for the present classification background. The constraints of thresholds in (6e) can be derived directly from (6c) and (6d). ■

Appendix II
Bounds Between Conditional Entropy and Bayesian Error in Binary Classifications

In the study of relations between mutual information (I) and Bayesian error (E), two important studies are reported on the LB by Fano [35] and the UB by Kovalevskij [36] in the forms of

$$\text{LB: } E \geq \frac{H(T) - I(T;Y) - H(E)}{log_2(m-1)} = \frac{H(T|Y) - H(E)}{log_2(m-1)} \tag{B1}$$

$$\text{UB: } E \leq \frac{H(T) - I(T;Y)}{2} = \frac{H(T|Y)}{2} \tag{B2}$$

where m is the total number of classes in T, $H(E)$ is the binary Shannon entropy, and $H(T|Y)$ is called conditional entropy, which can be derived from a general relation [2]

$$I(T;Y) = I(Y;T) = H(T) - H(T|Y) = H(Y) - H(Y|T). \tag{B3}$$

In binary classifications ($m = 2$), a tighter Fano's bound in [37] and [38] is adopted. With the rationals of Bayesian error, we suggest the tighter UB and LB in the following:

$$\text{Modified LB: } H(E) \geq H(T|Y), \text{ and } 0 \leq E \tag{B4}$$

$$\text{Modified UB: } E \leq \min\left(p(t_1), p(t_2), \frac{H(T|Y)}{2}\right). \tag{B5}$$

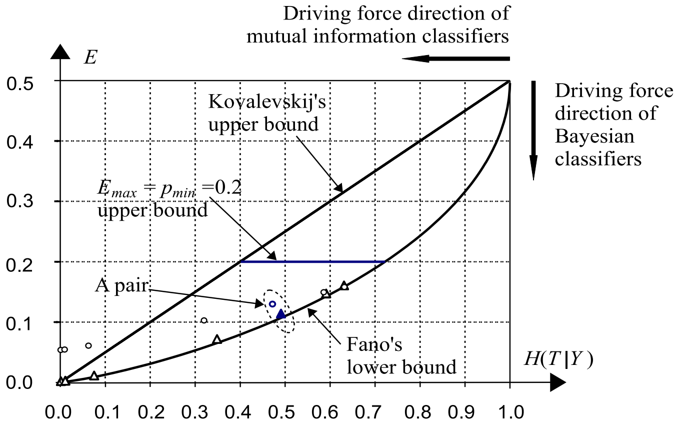


Fig. B.1. Bounds between conditional entropy $H(T|Y)$ and Bayesian error in binary classifications. Triangles and circles: the data in Table V from Bayesian and mutual-information classifiers, respectively. An UB of the Bayesian error exists, say, $E_{\max} = 0.2$ for the filled triangle.

Fig. B.1 shows the bounds in binary classifications, which is different from an E versus $I(T;Y)$ plot in [38]. From an equivalent relation in classifications [26]

$$\max\ I(T;Y) \leftrightarrow \min\ H(T|Y) \tag{B6}$$

the variable of $H(T|Y)$ is selected, because the two bounds are described directly by it. Triangles and circles shown in Fig. B.1 represent the paired data in Table IV from Bayesian and mutual-information classifiers, respectively. They clearly demonstrate the specific forms in their positions within the same pairs. The circle position is either coincident or up and/or left to its counterpart. These forms are attributed to different directions of driving force from two types of classifiers. One is for min E and the other for min $H(T|Y)$.

Numerical results present that the Fano's LB is effective for all classifiers, including mutual-information classifiers. The UBs, however, become invalid for mutual information classifiers. An UB of $E_{\max}$ ($= p_{\min}$) exists according to Theorem 3, which can improve Kovalevskij's one [36] with tightness in some range of $H(T|Y)$. When $E_{\max}$ decreases as shown in Table IV, the UB of the Bayesian error will become closer to its associated data.

Acknowledgment

The author would like to thank the assistance of Dr. Y.-J. Qu and C. Ocier for the text preparation. He would also like to thank the anonymous reviewers and the editors for their valuable comments and suggestions.

References

[1] S. R. Kulkarni, G. Lugosi, and S. S. Venkatesh, "Learning pattern classification—A survey," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2178–2206, Oct. 1998.

[2] R. O. Duda, P. E. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York, NY, USA: Wiley, 2001.

[3] Q. Yang and X. Wu, "10 challenging problems in data mining research," *Int. J. Inf. Technol. Decision Making*, vol. 5, no. 4, pp. 597–604, Nov. 2006.

[4] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 1, pp. 63–77, Jan. 2006.

[5] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[6] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[7] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial to the special issue on learning from imbalanced data sets," *ACM SIGKDD Explorations*, vol. 6, no. 1, pp. 1–6, Jun. 2004.

[8] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. 17th IJCAI*, 2001, pp. 973–978.

[9] C. Chow, "An optimum character recognition system using decision functions," *IRE Trans. Electron. Comput.*, vol. 6, no. 4, pp. 247–254, Dec. 1957.

[10] C. Chow, "On optimum recognition error and reject tradeoff," *IEEE Trans. Inf. Theory*, vol. 16, no. 1, pp. 41–46, Jan. 1970.

[11] A. Guerrero-Curieses, J. Cid-Sueiro, R. Alaiz-Rodríguez, and A. R. Figueiras-Vidal, "Local estimation of posterior class probabilities to minimize classification errors," *IEEE Trans. Neural Netw.*, vol. 15, no. 2, pp. 309–317, Mar. 2004.

[12] F. Tortorella, "A ROC-based reject rule for dichotomizers," *Pattern Recognit. Lett.*, vol. 26, no. 2, pp. 167–180, Jan. 2005.

[13] C. C. Friedel, U. Ruckert, and S. Kramer, "Cost curves for abstaining classifiers," in *Proc. ICML Workshop ROC Anal. ROCML*, 2006, pp. 33–40.

[14] T. Pietraszek, "On the use of ROC analysis for the optimization of abstaining classifiers," *Mach. Learn.*, vol. 68, no. 2, pp. 137–169, Aug. 2007.

[15] S. Vanderlooy, I. G. Sprinkhuizen-Kuyper, E. N. Smirnov, and H. Jaap van den Herik, "The ROC isometrics approach to construct reliable classifiers," *Intell. Data Anal.*, vol. 13, no. 1, pp. 3–37, Jan. 2009.

[16] E. Grall-Maës and P. Beauseroy, "Optimal decision rule with class-selective rejection and performance constraints," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 2073–2082, Nov. 2009.

[17] B. Dubuisson and M. Masson, "A statistical decision rule with incomplete knowledge about classes," *Pattern Recognit.*, vol. 26, no. 1, pp. 155–165, Jan. 1993.

[18] T. M. Ha, "The optimum class-selective rejection rule," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 6, pp. 608–615, Jun. 1997.

[19] D. J. C. Mackay, *Information Theory, Inference, and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[20] M. Adankon, M. Cheriet, and A. Biem, "Semisupervised learning using Bayesian interpretation: Application to LS-SVM," *IEEE Trans. Neural Netw.*, vol. 22, no. 4, pp. 513–524, Apr. 2011.

[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006.

[22] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[23] T. O. Kvalseth, "Entropy and correlation: Some comments," *IEEE Trans. Syst., Man, Cybern.*, vol. 17, no. 3, pp. 517–519, May 1987.

[24] I. Kononenko and I. Bratko, "Information-based evaluation criterion for classifier's performance," *Mach. Learn.*, vol. 6, no. 1, pp. 67–80, 1991.

[25] J. C. Principe, J. W. Fisher, and D. Xu, "Information theoretic learning," in *Unsupervised Adaptive Filtering*. New York, NY, USA: Wiley, 2000, pp. 265–319.

[26] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. New York, NY, USA: Springer-Verlag, 2010.

[27] Y. Wang and B.-G. Hu, "Derivations of normalized mutual information in binary classifications," in *Proc. 6th Int. Conf. Fuzzy Syst. Knowl. Discovery*, Aug. 2009, pp. 155–163.

[28] B.-G. Hu and Y. Wang, "Evaluation criteria based on mutual information for classifications including rejected class," *Acta Autom. Sinica*, vol. 34, no. 11, pp. 1396–1403, 2008.

[29] B.-G. Hu, R. He, and X.-T. Yuan, "Information-theoretic measures for objective evaluation of classifications," *Acta Autom. Sinica*, vol. 38, no. 7, pp. 1160–1173, 2012.

[30] C. M. Santos-Perira and A. M. Pires, "On optimal reject rules and ROC curves," *Pattern Recognit. Lett.*, vol. 26, no. 7, pp. 943–952, 2005.

[31] S. H. Yang, B.-G. Hu, and P.-H. Counede, "Structural identifiability of generalized constraints neural network models for nonlinear regression," *Neurocomputing*, vol. 72, nos. 1–3, pp. 392–400, Dec. 2008.

[32] B.-G. Hu, H.-B. Qu, Y. Wang, and S.-H. Yang, "A generalized constraint neural networks model: Associating partially known relationships for nonlinear regressions," *Inf. Sci.*, vol. 179, no. 12, pp. 1929–1943, May 2009.

[33] S.-H. Yang and B.-G. Hu, "Discriminative feature selection by nonparametric Bayes error minimization," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 8, pp. 1422–1434, Aug. 2012.

[34] B. Liu, "A survey of entropy of fuzzy variables," *J. Uncertain Syst.*, vol. 1, no. 1, pp. 4–13, 2007.

[35] R. M. Fano, *Transmission of Information: A Statistical Theory of Communication*. Cambridge, MA, USA: MIT Press, 1961.

[36] V. A. Kovalevskij, "The problem of character recognition from the point of view of mathematical statistics," in *Proc. Character Readers Pattern Recognit.*, 1968, pp. 3–30.

[37] I. Vajda and J. Zvárová, "On generalized entropies, Bayesian decisions and statistical diversity," *Kybernetika*, vol. 43, no. 5, pp. 675–696, 2007.

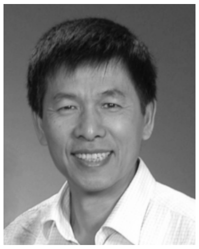
[38] J. W. Fisher, M. Siracusa, and K. Tieu, "Estimation of signal information content for classification," in *Proc. IEEE DSP Workshop*, 2009, pp. 353–358.

[39] G. M. Weiss and F. Provost, "The effect of class distribution on classifier learning," Dept. Comput. Sci., Rutgers Univ., Newark, NJ, USA, Tech. Rep. ML-TR 43, 2001.

[40] X.-W. Zhang and B.-G. Hu, "Learning in the class imbalance problem when costs are unknown for errors and rejects," in *Proc. IEEE 12th ICDM Workshop Cost Sensitive Data Mining*, Dec. 2012, pp. 194–201.

[41] M. Lin, K. Tang, and X. Yao, "Dynamic sampling approach to training neural networks for multiclass imbalance classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 4, pp. 647–660, Apr. 2013.

[42] C. L. Castro and A. P. Braga, "Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 6, pp. 888–899, Jun. 2013.



Bao-Gang Hu (M'94–SM'99) received the M.Sc. degree from the University of Science and Technology, Beijing, China, in 1983, and the Ph.D. degree from McMaster University, Hamilton, ON, Canada, in 1993, both in mechanical engineering.

He is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. From 2000 to 2005, he was the Chinese Director of Chinese-French Joint Laboratory for Computer Science, Control and Applied Mathematics (LIAMA). His current research interests include pattern recognition and plant growth modeling.