MULTICLASS CLASSIFICATION BASED ON BINARY CLASSIFIERS: ON CODING MATRIX DESIGN, RELIABILITY AND MAXIMUM NUMBER OF CLASSES

Sviatoslav Voloshynovskiy, Oleksiy Koval, Fokko Beekhof and Taras Holotyak

University of Geneva
Department of Computer Science
7 route de Drize, CH 1227, Geneva, Switzerland

ABSTRACT

In this paper, we consider the multiclass classification problem based on independent set of binary classifiers. Each binary classifier represents the output of quantized projection of training data onto a randomly generated orthonormal basis vector thus producing a binary label. The ensemble of all binary labels forms an analogue of a coding matrix. The properties of such kind of matrices and their impact on the maximum number of uniquely distinguishable classes are analyzed in this paper from an information-theoretic point of view. We also consider a concept of reliability for such kind of coding matrix generation that can be an alternative way for other adaptive training techniques and investigate the impact on the bit error probability. We demonstrate that it is equivalent to the considered random coding matrix without any bit reliability information in terms of recognition rate.

1. INTRODUCTION

In this paper we will address the multiclass categorization problem in a Machine Learning formulation requiring the assignments of labels to instances that belong to a finite set of classes ($M > 2$). While multiclass versions of most classification algorithms exist (e.g., [1]), they tend to be complex [2]. Therefore, a more common approach is to construct the multiclass classifier by combining the outputs of several binary classifiers [3, 4] that also extends to *error correcting output codes* (ECOC).

The ECOC framework consists of two main steps: a *coding* step, where the codeword or some representation of entry is assigned to a row of a coding matrix, and a *decoding* step, where a given observation is mapped into the most similar codeword of coding matrix. There are many methods of coding matrix design based on the predefined set of codewords that follow different heuristics with the overall idea to maximize the inter-codeword Hamming distances that is believed to correspond to the most robust coding matrix design in terms of classification accuracy. However, these predefined coding matrices are problem-independent and can not cover a broad class of varying models. The design of an optimal decoder minimizing the overall misclassification error probability is also mainly accomplished based on the minimum Hamming distance decoder that is a form of hard decoding. Although several score-based decoding rules (e.g., loss-based and loss-weighted decoding) attempt to consider the effect of binary classification reliability in the overall fusion rule, the theoretically justified probabilistic fusion rules are still missing. Despite several recent remarkable exceptions [5, 6, 7], these problems are little studied and the problem of joint coding matrix design and probabilistic decoding maximizing the number of uniquely recognizable classes is of great practical interest.

The contact author is S. Voloshynovskiy (email: svolos@unige.ch). http://sip.unige.ch. This paper was partially supported by SNF projects 200021-111643 and 200021-1119770 and Swiss IM2 project.

2. PROBLEM FORMULATION

In this paper we will follow the information-theoretic machine learning approach thus providing the link with the coding theory for optimal joint coding matrix and decoder design and estimation of the maximum number of uniquely distinguishable classes.

Assuming that the data are independent or weakly dependent and can be treated as almost identically distributed, one can use the definition of *information density*:

$$I_N = \frac{1}{N} \log_2 \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}, \quad (1)$$

where $\mathbf{x}$ is the template for the learning set and $\mathbf{y}$ is a template of the data to be classified, N is the template length, and $p(\mathbf{x}, \mathbf{y})$, $p(\mathbf{x})$ and $p(\mathbf{y})$ are joint probability density of $\mathbf{X}$ and $\mathbf{Y}$ and their marginals, respectively. When the template distributions are known, a so-called *recognition* or *identification capacity* [8] can be used:

$$\bar{I}(X;Y) = \lim_{N\to\infty} E[I_N], \quad (2)$$

provided that the limit is well defined and the expectation is taken with respect to the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$

that reduces to the Kullback-Leibler Distance (KLD) between $p(\mathbf{x}, \mathbf{y})$ and $p(\mathbf{x})p(\mathbf{y})$. This also corresponds to the Bayesian multiclass classifier minimizing the average probability of misclassification and is invariant under linear invertible transformations. In this case, the maximum number of classes that can be recognized with vanishing probability of error under the above conditions is limited as [9]:

$$M \leq 2^{N\bar{I}(X;Y)}. \quad (3)$$

For the case of i.i.d. Gaussian data $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I}_N)$ and the memoryless additive white Gaussian model of interaction $\mathbf{y} = \mathbf{x} + \mathbf{z}$ with $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma_Z^2 \mathbf{I}_N)$, the recognition capacity is readily found as:

$$\bar{I}(X;Y) = \frac{1}{2}\log_2 \frac{1}{1-\rho_{XY}^2} = \frac{1}{2}\log_2\left(1 + \frac{\sigma_X^2}{\sigma_Z^2}\right), \quad (4)$$

where $\rho_{XY}^2 = \frac{\sigma_X^2}{\sigma_X^2+\sigma_Z^2}$ is a squared correlation coefficient (SCC) between X and Y. The results can be extended to a more general model of interaction with training model $p(\mathbf{v}|\mathbf{x})$ and observation model $p(\mathbf{y}|\mathbf{x})$. For the i.i.d. Gaussian case with the training data model $\mathbf{v} = \mathbf{x} + \mathbf{z}_t$ and observation model $\mathbf{y} = \mathbf{x} + \mathbf{z}_r$ with $\mathbf{Z}_t \sim \mathcal{N}(\mathbf{0}, \sigma_{Z_t}^2 \mathbf{I}_N)$ and $\mathbf{Z}_r \sim \mathcal{N}(\mathbf{0}, \sigma_{Z_r}^2 \mathbf{I}_N)$, the recognition capacity is:

$$\bar{I}(V;Y) = \frac{1}{2}\log_2 \frac{1}{1-\rho_{VY}^2}, \quad (5)$$

where $\rho_{VY}^2 = \frac{\sigma_X^4}{(\sigma_X^2+\sigma_{Z_t}^2)(\sigma_X^2+\sigma_{Z_r}^2)}$ is the SCC between V and Y that transforms $\rho_{VY}^2 = \rho_{XY}^2$ for $\mathbf{v} = \mathbf{x}$.

3. PROPOSED APPROACH

Instead of following the above discussed construction of coding matrix and training the classifiers, we will consider a scheme that targets maximization of the number of correctly distinguishable classes based on the binary classification. We will demonstrate that the structure of the coding matrix that corresponds to the above objective and simultaneously maximizes minimum distance is obtained directly from the training stage by mapping each vector of a training set into a row of the coding matrix.

Without loss of generality, we will assume at this stage that we have a mapper/encoding function $f(.)$ that maps the training set and observation entries as $f : \mathcal{X}^N \rightarrow \mathcal{B}_{\mathbf{x}}^L$, $\mathcal{B}_{\mathbf{x}} \in \{0,1\}$, and $f : \mathcal{Y}^N \rightarrow \mathcal{B}_{\mathbf{y}}^L$, $\mathcal{B}_{\mathbf{y}} \in \{0,1\}$, respectively. Therefore, the link between the binary representation $\mathbf{b}_{\mathbf{x}}$ of vector $\mathbf{x}$ and its noisy counterpart $\mathbf{b}_{\mathbf{y}}$ of vector $\mathbf{y}$ is defined according to a *binary symmetric channel* (BSC) model [9]. We assume that noise in the direct domain might cause a bit flipping in the binary domain with a certain average probability $\bar{P}_b$. The corresponding maximum number of recognizable classes (3) can be readily found as [9]:

$$M_b \leq 2^{L\bar{I}(B_{\mathbf{x}};B_{\mathbf{y}})}. \quad (6)$$

Extending the mutual information between binary representations or classifiers outputs, one obtains:

$$\bar{I}(B_{\mathbf{x}}; B_{\mathbf{y}}) = H(B_{\mathbf{x}}) - H(B_{\mathbf{x}}|B_{\mathbf{y}}). \quad (7)$$

It can be immediately noticed that to maximize the M_b, one needs to maximize $I(B_{\mathbf{x}}; B_{\mathbf{y}})$ for a given L that can be achieved by: (a) maximization of $H(B_{\mathbf{x}})$ and (b) minimization of $H(B_{\mathbf{x}}|B_{\mathbf{y}})$. In the considered binary case, the maximum value of term $H(B_{\mathbf{x}})$ is 1, that can be achieved for equiprobable independent data, i.e., $\Pr(0) = \Pr(1) = 0.5$. This suggests that the multi-class rate maximization coding matrix should have equiprobable independent binary entries that is known as a random coding matrix.

The second term $H(B_{\mathbf{x}}|B_{\mathbf{y}})$ is defined by the average error probability of binary classification $\bar{P}_b$ and $H(B_{\mathbf{x}}|B_{\mathbf{y}}) = H_2(\bar{P}_b) = -\bar{P}_b \log_2 \bar{P}_b - (1-\bar{P}_b)\log_2(1-\bar{P}_b)$ that is the binary entropy. In the considered setup it is not possible to control $\bar{P}_b$. Therefore, we will consider an alternative design where $\bar{P}_b$ can be considerably reduced due to basis adaptation based on decision reliability information. At the same time, we will demonstrate that this decrease of bit error probability comes in price of increased size of coding matrix that equivalently reduces the recognition rate. In fact, we will demonstrate that these two approaches are equivalent in terms of recognition rate and simply represent different designs of coding matrices. The only increase of recognition rate can be achieved due to the change of fusion rule based on reliability information for a fixed size of the random coding matrix.

3.1. Design of coding matrix

According to the above analysis the coding matrix should maximize $H(B_{\mathbf{x}})$. Simultaneously, we have assumed the existence of a generic encoding function $f(.)$ that maps the real-data entries into binary representations stored in the coding matrix. To achieve the maximum of $H(B_{\mathbf{x}}) = 1$, $B_{\mathbf{x}}$ should be equiprobable and independent. In this section, we will consider a possible design of such kind of encoding function based on random projections and binarization.

The random projections are considered as a dimensionality reduction step and are performed as:

$$\tilde{\mathbf{x}} = \mathbf{W}\mathbf{x}, \quad (8)$$

where $\mathbf{x} \in \mathbb{R}^N$, $\tilde{\mathbf{x}} \in \mathbb{R}^L$, $\mathbf{W} \in \mathbb{R}^{L \times N}$ and $L \leq N$ and $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_L)^T$ consists of a set of projection basis vectors $\mathbf{w}_i \in \mathbb{R}^N$ with $1 \leq i \leq L$. Instead of following a particular consideration of mapping $\mathbf{W}$, we will assume that $\mathbf{W}$ is a random matrix. The matrix $\mathbf{W}$ has elements $w_{i,j}$ that are generated from some specified distribution. An $L \times N$ random matrix $\mathbf{W}$ whose entries $w_{i,j}$ are independent realizations of Gaussian random variables $W_{i,j} \sim \mathcal{N}(0, \frac{1}{N})$ is of a particular interest for our study.

In this case, such a matrix can be considered as an almost *orthoprojector*, for which $\mathbf{W}\mathbf{W}^T \approx \mathbf{I}_L$. [1]

The second step uses labeling or Grey codes to ensure closeness of labels for close vectors. Such kind of labeling is known as *soft hashing*. When only the most significant bit of the Grey code is used, it is known as binary or *hard hashing*.

The most simple quantization or binarization of extracted features is known as *sign random projections*:

$$b_{\mathbf{x}_i} = sign(\mathbf{w}_i^T \mathbf{x}), \quad (9)$$

where $b_{\mathbf{x}_i} \in \{0, 1\}$, with $1 \leq i \leq L$ and $sign(a) = 1$, if $a \geq 0$ and 0, otherwise. The vector $\mathbf{b}_\mathbf{x} \in \{0, 1\}^L$ computed for all projections represents a binary label of class computed from the vector $\mathbf{x}$. The ensemble of all binary labels $\mathbf{b}_\mathbf{x}(m)$ with $1 \leq m \leq M$ forms a coding matrix.

It can be readily validated that due to the independence of projections the results of projections will be independent and almost equiprobable that satisfies the necessary conditions of entropy maximization.

At the same time, one can notice that the binary labels are deduced directly from the training data and stored in the coding matrix thus avoiding the additional stage of matching binary label deduced from the training vector with the closest row of the coding matrix as it is done for the methods discussed in the introduction.

3.2. Minimization of average error probability for binary classifiers: reliability function

The bit error probability indicates the mismatch of signs between $\tilde{x}_i$ and $\tilde{y}_i$, i.e., $\Pr[sign(\tilde{x}_i) \neq sign(\tilde{y}_i)]$. For a given $\mathbf{x}$ and $\mathbf{w}_i$, one can find the probability of bit error as:

$$P_{b|\tilde{x}_i} = \frac{1}{2}(\Pr[\tilde{Y}_i \geq 0 | \tilde{X}_i < 0] + \Pr[\tilde{Y}_i < 0 | \tilde{X}_i \geq 0]), \quad (10)$$

or by symmetry as:

$$P_{b|\tilde{x}_i} = \Pr[\tilde{Y}_i < 0 | \tilde{X}_i \geq 0]. \quad (11)$$

For a given $\tilde{x}_i$ and Gaussian noise[2], the distribution of the projected vector is $\tilde{Y}_i \sim \mathcal{N}(\tilde{x}_i, \sigma_Z^2 \mathbf{w}_i^T \mathbf{w}_i)$ that reduces to $\tilde{Y}_i \sim \mathcal{N}(\tilde{x}_i, \sigma_Z^2)$ for the orthoprojection case ($\mathbf{w}_i^T \mathbf{w}_i = 1$) and:

$$P_{b|\tilde{x}_i} = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi\sigma_Z^2}} e^{\frac{-(\tilde{y}_i - \tilde{x}_i)^2}{2\sigma_Z^2}} d\tilde{y}_i = Q\left(\frac{\tilde{x}_i}{\sigma_Z}\right). \quad (12)$$

The origin of $P_{b|\tilde{x}_i}$ for a given configuration of $\mathbf{x}$ and $\mathbf{w}_i$ is shown in Figure 1. The vector $\mathbf{x}$ forms the angle θ_{XW_i} with the basis vector $\mathbf{w}_i$ and the projection results into the scalar value $\tilde{x}_i$. The closer angle θ_{XW_i} to $\pi/2$, the smaller value $\tilde{x}_i$. This leads to the larger probability that the sign of $\tilde{y}_i$ will be different from the sign of $\tilde{x}_i$ that is shown by the gray area under the curve of $p(\tilde{y}_i)$. One can immediately note that since the projections are generated at random there is generally no guaranty that two vectors can be collinear. However, at the same time some of the projections might form angles with $\mathbf{x}$ that deviate from $\pi/2$ thus leading to a smaller probability of binary classification error. This observation makes it possible to assume that some projections can be more preferable than others and the equation (12) can be a good measure of bit *reliability*. We will denote the reliability of the ith bit computed from $\mathbf{x}$ for projection $\mathbf{w}_i$ as $R_\mathbf{x}(i) = 1 - P_{b|\tilde{x}_i} = 1 - Q\left(\frac{\tilde{x}_i}{\sigma_Z}\right)$. Obviously, the larger the value of $\tilde{x}_i$ is obtained in the result of projection, the closer $R_\mathbf{x}(i)$ is to 1. The above analysis refers

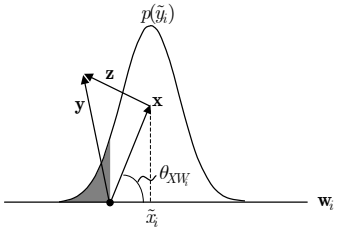

Fig. 1. The bit error probability for a given $\mathbf{x}$ and some $\mathbf{w}_i$.

to only one realization $\mathbf{x}$. Since $\mathbf{X}$ is a random vector following some distribution $p(\mathbf{x})$, one should find the average probability of error for all possible realizations. Assuming $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I}_N)$, the statistics of data in the projection domain are $\tilde{X}_i \sim \mathcal{N}(0, \sigma_X^2)$ (Figure2,a) and the average bit error probability is:

$$\bar{P}_b = 2 \int_0^\infty P_{b|\tilde{x}_i} p(\tilde{x}_i) d\tilde{x}_i \quad (13)$$

$$= 2 \int_0^\infty Q\left(\frac{\tilde{x}_i}{\sigma_Z}\right) \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{\frac{-\tilde{x}_i^2}{2\sigma_X^2}} d\tilde{x}_i = \frac{1}{\pi} \arccos(\rho_{XY}). \quad (14)$$

Remarkably, the average probability of error depends on the correlation coefficient between the direct domain data and is determined by the channel and source statistics. It can be easily verified for the more general model that the average bit error probability (13) is:

$$\bar{P}_b = \pi^{-1} \arccos(\rho_{VY}), \quad (15)$$

that coincides with (13) for the noiseless training case. Obviously, this sort of ambiguity during training causes an additional increase in probability since $\rho_{XY} \geq \rho_{VY}$, this will be demonstrated by the results of computer simulation.

It is also important to note that all possible values $\tilde{x}_i$ in (13) originating from both "unreliable", i.e., values are close to zero, and "reliable", i.e., values are far away from zero, projections are taken into account with the same weight to

[1] Otherwise, one can apply special orthogonalization techniques to ensure perfect orthogonality.

[2] In the case of assumed Gaussian random basis vectors $\mathbf{w}_i$ any distribution will be mapped into Gaussian one for both entry and noise data.

form the resulting binary vector $\mathbf{b_x}$ and corresponding coding matrix. Obviously, for a given set of training data one can always find a set of vectors $\mathbf{w}_i$, $1 \leq i \leq L$ that minimizes the overall bit error probability. However, keeping in mind the facts that (a) the number of classes might be of order of millions; and (b) it can be constantly updated; such an optimization problem looks highly unfeasible.

Therefore, in the scope of this paper we will consider another approach when one generates the *overcomplete set of projections* J and selects among them only those L projections that are the largest in the absolute magnitude, if a fixed number of bits is requested, or those that are higher then a certain threshold $T_{\tilde{x}}$ for a given $\mathbf{x}$.

In case of the above thresholding approach (Figure 2,b), the corresponding average probability of bit error is:

$$\bar{P}_{b_T} = \frac{1}{\int_{T_{\tilde{x}}}^{\infty} p(\tilde{x}_i)d\tilde{x}_i} \int_{T_{\tilde{x}}}^{\infty} P_{b|\tilde{x}_i} p(\tilde{x}_i) d\tilde{x}_i \quad (16)$$

$$= Q^{-1}\left(\frac{T_{\tilde{x}}}{\sigma_X}\right) \int_{T_{\tilde{x}}}^{\infty} Q\left(\frac{\tilde{x}_i}{\sigma_{Z_r}}\right) \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{\frac{-\tilde{x}_i^2}{2\sigma_X^2}} d\tilde{x}_i, \quad (17)$$

where the multiplier is the normalization constant corresponding to the fraction of distribution behind the threshold.

The practical application of this approach is facing three main concerns: (a) which number of overcomplete projections J is needed for any $\mathbf{x}$ to guarantee the necessary L?; (b) what is a possible gain in $\bar{P}_{b_T}$ versus $\bar{P}_b$?; (c) what is the impact on $\bar{P}_b$ of a mismatch between the reliable projections extracted from $\mathbf{v}$ at the training stage and those extracted from $\mathbf{y}$ at the classification stage? We will address the issues (a) and (b) in this section to demonstrate the feasibility of the proposed approach and possible increase in the classification accuracy and investigate the remaining issue (c) in the experimental part of the paper. At the same time, one should take into account the increase of the coding matrix size L to store the information about reliable projections that might affect the achievable recognition rate.

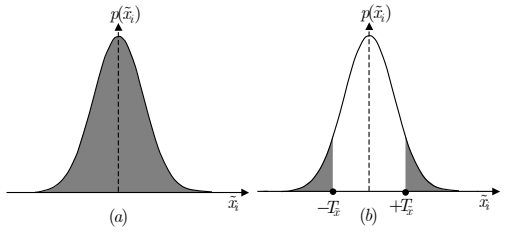

Fig. 2. Bit error reliability framework: (a) all values $\tilde{x}_i$ are taken into account and (b) only the most reliable values are taken for the basis selection (to be normalized to 1).

It is easy to verify that the number of coefficients L of random variable $\tilde{X}_i$ following Gaussian distribution and exceeding the threshold $T_{\tilde{x}}$ in J projections satisfies with high probability the following equation:

$$\Pr[L \geq \ell] = 1 - F_{B_X}(J, \ell, \Pr[\tilde{X}_i > T_{\tilde{x}}]), \quad (18)$$

where ℓ is the necessary number of reliable coefficients in the coding matrix (like 32, 64 or 128), $F_{B_X}(J, \ell, \Pr[\tilde{X}_i > T_{\tilde{x}}])$ designates binomial cumulative distribution function and $\Pr[\tilde{X}_i > T_{\tilde{x}}] = Q\left(\frac{T_{\tilde{x}}}{\sigma_X}\right)$. For practical applications, one can assume that to ensure the existence of desired ℓ with high probability $1-\epsilon$, the quantity $F_{B_X}(J, \ell, \Pr[\tilde{X}_i > T_{\tilde{x}}])$ should be bounded by a small ϵ that will be further assumed not to exceed 10^{-10}. This result is shown in Figure 3. Obviously, the larger the number of reliable bits is requested in the coding matrix, the more projections J should be generated for a given threshold. At the same time, the increase of the threshold leads to the exponential number of projections J. Although these numbers seem to be quite high, for example for $L = 64$ and $T_{\tilde{x}} = 2.5$, the needed J is about $2 \cdot 10^4$, this can be compared to the discrete Fourier transform of image of size 512×512 for the optimal feature selection out of about $2.6 \cdot 10^5$ transform coefficients. Therefore, this problem is computationally feasible. To answer the second

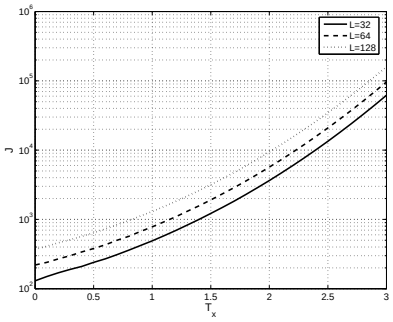

Fig. 3. The estimation of necessary number of projections J for the desired number of reliable bits in coding matrix for $\sigma_X^2 = 1$.

question about the possible gain in $\bar{P}_{b_T}$ versus $\bar{P}_b$, we will plot the corresponding results (13) and (16) for different $T_{\tilde{x}}$ as a function of signal-to-noise ratio (SNR) defined as $\text{SNR} = 10\log_{10}\frac{\sigma_X^2}{\sigma_{Z_r}^2}$. The results are shown in Figure 4. The proposed optimization strategy to the reliable projection selection clearly demonstrates a considerable increase in the accuracy of binary classifiers with respect to the blind projection selection. The results coincide for $T_{\tilde{x}} = 0$ that confirms the fact that all projections are blindly taken into account for the coding matrix generation.

3.3. Practical decoders

There are several possible practical implementations of the proposed framework. The first approach consists in generating J projections and selecting only the L most reliable ones thus producing a $M \times L$ coding matrix and a $M \times L'$

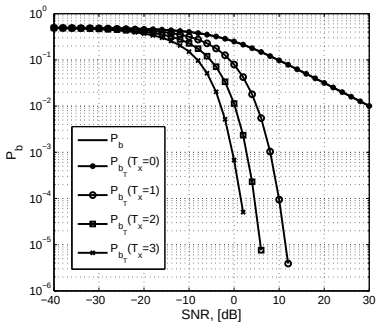

Fig. 4. The average bit error probability for blind and reliable projections selection for various thresholds.

matrix storing information about reliable components' positions. For one entry the reliable components' position vector consists of $\mathbf{P_x} = \mathbf{1}(\mathbf{R_x}) \in \{0,1\}^J$, with L non-zero components and $\mathbf{1}(.)$ denotes the indicator function taking value 1, if the the reliability function corresponds to the reliable bit and 0, otherwise. If the information is encoded encoded directly $L' = H(\mathbf{P_x}) = JH_2(P_{p_X})$, where P_{p_X} is the probability of 1 in the position vector. Another alternative consists in the fact that $\mathbf{P_x}$ and its counterpart $\mathbf{P_y}$ are correlated and one can use a distributed source coding technique based on binning [9] thus reducing the equivalent length to $L' = H(\mathbf{P_x}|\mathbf{P_y}) = JH_2(P_{p_X|p_Y})$, where $P_{p_X|p_Y}$ is the probability of bit mismatch between $\mathbf{P_x}$ and $\mathbf{P_y}$. Disregarding the storage format, the reliability based decoder filters out all unreliable bits in the distance metric preserving only L reliable bits:

$$\hat{m} = \arg \min_{1 \le m \le M} \sum_{j=1}^{J} d^H(b_{\mathbf{x}_j}(m), b_{\mathbf{y}_j}) P_{\mathbf{y}_j}, \tag{19}$$

where $\mathbf{b_x} \in \{-1,1\}^J$ and $\mathbf{b_y} \in \{-1,1\}^J$. We will show that indeed one can achieve a considerable decrease in the bit error probability in cost of the coding matrix size increase. We will demonstrate in the next section that the recognition rate of this approach is equivalent to those based on $M \times L$ blind random coding matrix proposed in the paper.

4. RESULTS OF COMPUTER SIMULATION AND CONCLUSIONS

In this section, we will first demonstrate the maximum number of uniquely distinguishable classes in terms of recognition rates computed for the best achievable case of direct domain classification $\bar{I}(X;Y)$ according to (4) and based on a set of L equivalent binary classifiers with the blind selection of projections $\bar{I}(B_{\mathbf{x}};B_{\mathbf{y}})$ and reliability based projection selection $\tilde{I}(B_{\mathbf{x}};B_{\mathbf{y}})$. In the second part of modeling, we present empirical bit error probabilities for different classification strategies and highlight the mismatch in training and recognition data models on the average probability of bit error.

The recognition rates for the above classifiers are shown in Figure 5 for the dimensionality reduction factor $L/N = 32/1024$. All results are obtained by simulation of 100 class realizations, 100 projection matrices and 100 noise realizations. The recognition rate of the optimal Bayesian multiclass classifier denoted as "Direct: dim. N" with the data dimensionality N represents the best achievable limit under the completely known distributions. The dimensionality reduction based on random projections leads to decrease of this rate by a factor L/N that is denoted as "Direct: dim. L", while all classifiers based on the set of binary classifiers achieve their limit equal to 1 at high SNR. We tested two classes of multiclass classifiers based on binary classifiers, i.e., blind and reliability-based, for two models of training data $(\mathbf{x},\mathbf{y})$ and $(\mathbf{v},\mathbf{y})$. The number of reliable projections was varying to keep the overall coding matrix and position information matrix length equal to L based on $J = 1500$ random projections for the reliability based binary classification. The proposed blind classifier achievable recognition rate coincides with the one of reliability based classifier. It should be also pointed out that the Bayesian classifier requires the knowledge of actual priors models while the proposed classifiers are based on the binary model and Gaussian statistics of reliability function in the discussed projected domain. The rate for noisy training data $\mathbf{v}$ based on the binary classifiers represent the lower bound in the performed modeling. It is interesting to emphasize that the rates for the reliability based classifiers with the reliability estimation based on $\mathbf{R_x}$ and $\mathbf{R_y}$ coincide.

The empirical average bit error probabilities for different classification strategies are shown in Figure 6. To investigate the impact of the above mismatch on the training data, we performed tests for three different cases when the reliability function is computed from the original training data $\mathbf{x}$, training noisy data $\mathbf{v}$ and directly observation data $\mathbf{y}$. Obviously, two first cases are just presented for illustration purposes since they assume that the class index m is known at the classifier for the optimal projection basis selection that is not the case for the practical systems. Nevertheless, it is interesting to note that the impact of mismatch in the accuracy of optimal projections selection based on the reliability functions $\mathbf{R_x}$ and $\mathbf{R_y}$ and $\mathbf{R_v}$ and $\mathbf{R_y}$ in the corresponding models $(\mathbf{x},\mathbf{y})$ and $(\mathbf{v},\mathbf{y})$ is negligible and the curves practically coincide. The reason for that is explained by a small number of wrongly identified optimal projections directly from noisy data with respect to two training models. The mismatch in the number of correctly estimated projections for $J = 1500$ random projections based on the distance between the indicator functions $\mathbf{P_y}$ vs $\mathbf{P_x}$, and $\mathbf{P_y}$ vs $\mathbf{P_v}$ is shown in Figure 7. This repre-

sents a small (less than 5%) mismatch that does not have any practically significant impact on the average bit error and demonstrates the high robustness of the proposed method to training and model errors. Finally, it considerably outperfroms the blind binary classifier for both models $(\mathbf{x}, \mathbf{y})$ and $(\mathbf{v}, \mathbf{y})$ in terms of bit error rate. However, this advantage is equilized for the achieveable recognition rate due to the storage of reliable bits' position information. The

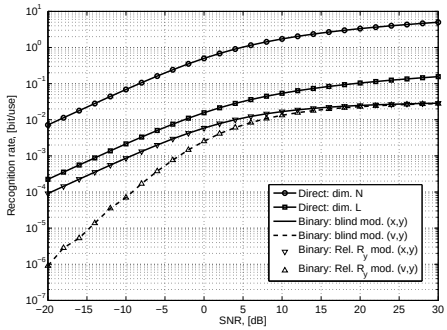

Fig. 5. The maximum number of uniquely distinguishable classes in terms of recognition rates for the direct and random projections domains multiclass classifier and classifiers based on a set of binary classifiers with blind and reliability based projections selection.

proposed approach also considerably outperforms in terms of the classification rate both the one-vs-one coding matrix design $R_{\text{o-v-o}} = 1/N \log_2 L/L$ and the one-vs-all design $R_{\text{o-v-a}} = 1/N \log_2 L/((L^2 - L)/2)$. Therefore, being built into the multiclass classification framework the proposed methods automatically suggest the optimal design of coding matrix that maximizes the number of uniquely recognizable classes, do not need retraining of binary classifiers for the addition of new classes and are robust to the training and model errors. The proposed approach based on bit reliability can be extended to create new more powerful fusion rules based on real values of the projected coefficients mapped into the corresponding weights and to reduce the complexity of the multiclass classification with large M.

5. REFERENCES

[1] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2001.

[2] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multi-class support vector machines," *IEEE Trans. on Neural Networks*, vol. 2, pp. 415–425, 2002.

[3] T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *Journal of Artificial Intelligence Research*, vol. 2, pp. 263–286, 1995.

[4] E. L. Allwein, R. E. Schapire, Y. Singer, and P. Kaelbling, "Reducing multiclass to binary: A unifying approach for margin classifiers," *Journal of Machine Learning Research*, vol. 1, pp. 113–141, 2000.

[5] S. Escalera, O. Pujol, and P. Radeva, "Loss-weighted decoding for error-correcting output coding," in *3rd Int. Conf. on Computer Vision Theory and Applications*, Madeira, Portugal, 22 - 25 Jan. 2008, pp. 117–122.

[6] O. Dekel and Y. Singer, "Multiclass learning by probabilistic embeddings," in *In NIPS*, 2002, pp. 945–952.

[7] A. Passerini, M. Pontil, and P. Frasconi, "New results on error correcting output codes of kernel machines," *IEEE Trans. on Neural Networks*, vol. 1, pp. 45–54, 2004.

[8] J. A. OSullivan and N. A. Schmid, "Performance analysis of physical signature authentication," *IEEE Trans. on Info. Theory*, vol. 47, pp. 3034–3039, 2001.

[9] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley and Sons, New York, 1991.

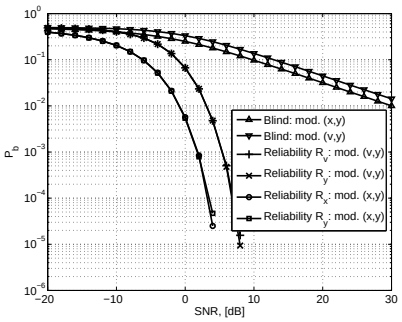

Fig. 6. Average bit error probability for blind and reliability based binary classifiers with different training strategies.

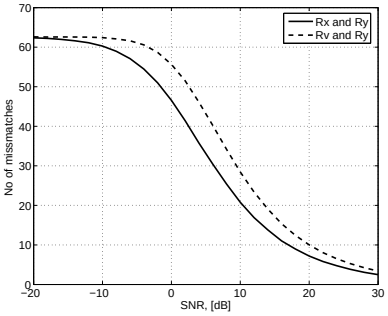

Fig. 7. Number of incorrectly identified optimal projections in pairs $\mathbf{P_y}$-$\mathbf{P_x}$ and $\mathbf{P_y}$-$\mathbf{P_v}$ for $J = 1500$.