MULTI-CLASS CLASSIFIERS BASED ON BINARY CLASSIFIERS: PERFORMANCE, EFFICIENCY, AND MINIMUM CODING MATRIX DISTANCES

Fokko Beekhof, Sviatoslav Voloshynovskiy, Oleksiy Koval and Taras Holotyak

7, route de Drize
CH-1227 Carouge (Geneva)
Switzerland

ABSTRACT

Using multiple binary classifiers is a popular way to construct multi-class classifiers. There exist several strategies to construct multi-class classifiers from binary classifiers. An important question is which strategy offers the highest probability of successful classification given the number of binary classifiers used. The first result presented in this work is a method to approximate how many classes can be distinguished using N binary classifiers in practical systems rather than theoretical setups. We come to the conclusion that in this formulation, all methods share the same performance limit, which is determined using the first result. The next question is what the smallest number of binary classifiers is that is needed to attain a given probability of success. To investigate this, we introduce the concept of efficiency, which is the ratio between the number bits needed to count the number of distinguishable classes and the number of bits used. The last contribution concerns the conclusion that methods should exist that are more efficient than those currently employed.

1. INTRODUCTION

In this paper we address the efficiency of multi-class classification strategies based on multiple binary classifiers. The efficiency in this context refers to the ability to attain a certain classification accuracy for a given number of classes and a certain probability of individual binary classifiers being in error when using a specific number of binary classifiers. This notion of efficiency is derived from the rate as defined in digital communications. The limits on what can be accomplished are readily determined in the general case using information-theoretic arguments [1], however, these arguments cannot be easily applied to finite numbers.

The contributions presented in this work are the following. First, we present an approximative method to determine the number of classes that can be handled using a finite number of classifiers. Second, we derive realistic limits on the capabilities of multi-class classification systems based on multiple binary classifiers using minimum Hamming-distance decoding. Third, we present a numerical validation of the approximative method, and subsequently show that the efficiency that said classification systems can accomplish is significantly lower than what one would suspect based on classic information-theoretic results. Finally, we show that, despite the fact that the limit on the achievable efficiency is much lower than for asymptotally large numbers of binary classifiers, existing strategies do not come close to attaining the efficiency that should practically be achievable. Conversely, this implies that much better classification strategies probably exist.

1.1. One-vs-All

One of the most common and intuitive ways of constructing a multi-class classifier is by creating a binary classifier for each class, that tests the hypothesis whether the associated class has been observed, or any of the others. The most famous of binary classifiers is probably SVM [2]. In the one-vs-all approach, ideally only one binary classifier should test positive during each classification. A problem is that it might be hard to select a particular class if a subset of classifiers gives conflicting answers, i.e. if at least two hypotheses test positive.

1.2. All-vs-All

Hastie and Tibshirani [3] proposed to use pairwise binary classifiers, the so-called all-vs-all approach. The output of the multiclass classifier is the class which has "won" the most pairwise classifications. Unfortunately, the number of classifiers is a quadratic function of the number of classes, which can cause scalability issues. In light of the central question of this work, which is how many binary classifiers are needed to successfully perform multi-class classification, such an abundant need for binary classifiers should be well justified.

1.3. ECOC

An extention to both approaches has been proposed by Dieterich and Bakiri [4], that is based on error-correction codes (ECC). Each sequence of bits produced by a set of binary classifiers is associated with codewords of an ECC during learning, with the intent of finding a coding matrix with superior properties. Linking classes to a pre-existing binary coding matrix is a delicate issue, it may for example be the case that although both the coding matrix derived from the classes directly and the ECC may contain well-separable entries, but there is no assurance that the separation boundaries of the two coincide. When the ECC uses more bits then there are binary classifiers, the efficiency is affected, and we should take this into account when evaluating the improvement in accuracy.

1.4. Performance: Link to Digital Communications

Allwein et al [5] constructed a framework for analyis of all three approaches and concluded that the one-vs-all strategy did not perform as well as the others. Rifkin and Klautau [6] then argued against this conclusion, and claimed that one-vs-all could perform just as well as the other schemes. As far as we know, this debate is still ongoing.

In this work we show that an evaluation of the accuracy of classification systems using multiple binary classifiers depends first on the relation between the number of classifiers N, the number of classes M, and the probability of error of the individual classifiers. Conversely, the probability of individual error can be used to determine a measure of efficiency called the classification rate R.

Based on arguments from communication theory, we arrive at the conclusion that in many circumstances it is not only true that one-vs-all should be able to perform equally well with proper binary classifiers, but even that other strategies probably exist that accomplish the same task with a significantly lower number of classifiers. Conversely, the same theory is known to show that when the individual classifiers are too often in error, classification cannot be performed correctly if the number of classifiers relative to the number of classes is too low.

Although these are known results in information- and communication theory, these generally hold for asymptotically large N. In this work, we propose a probabilistic method that can be used to establish the attainable rate for finite numbers and a chosen probability of classification error. The conclusion is that although the efficiency reduces significantly when only finite quantities are used, it is still possible to classify with the chosen probability of classification error using fewer binary classifiers than any of the discussed schemes.

2. ESTIMATING THE NUMBER OF DISTINGUISHABLE CLASSES

Let a "codeword" corresponding to a class be the binary sequence produced by N different binary classifiers when none of these classifiers is in error.

The number of distinguishable classes depends on the spread of codewords in the binary space and the number of bit-errors that appear due to errors of individual classifiers. In the following analysis, we assume that the outputs of the binary classifiers are i.i.d. and produce a one with probability p_1. Ideally, p_1 should be 0.5 because this implies the highest entropy, which can be shown to correspond to the highest efficiency. Furthermore, we assume that all classifiers are all independently and equally likely to be in error, and that the probability of an individual error is p_b. Under these assumptions, the number of classes that can be distinguished using binary classifiers can be determined.

Let us first consider the Hamming distance between two codewords. Then, two bits in the same position in both codewords are different with a probability $p_{neq} = 2p(1-p)$. Note that if $p = 0.5$, then also $p_{neq} = 0.5$. The Hamming distance D between two codewords is then Binomially distributed: $D \sim Bin(N, p_{neq})$. Thus we can state that:

$$\Pr[D >= d] = 1 - \Pr[D <= d + 1] \tag{1}$$

where $\Pr[D <= d + 1]$ can be determined using the cumulative density function of the Binomial distribution.

If the number of classes is M, a total of $\frac{M^2-M}{2}$ inter-codeword Hamming distances can be computed. In reality, these distances must be related, as can be understood from the fact that the smallest Hamming distance must forcibly be equal to zero if $M > 2^N$. The approximation presented here is obtained by considering the distances between each possible pair as independent events:

$$\Pr[\forall i, j D_{i,j} >= d] \approx (1 - \Pr[D <= d+1])^{\frac{M^2-M}{2}}. \tag{2}$$

It is important to notice that a particular minimum distance d_{min} is a probabilistic event in both the original and the approximate formulations, and, at all times, it is possible that d_{min} is zero with some probability. Zero is therefore a correct lower bound for d_{min}, but not a very useful one because it implies that there is no separability between certain classes. In order to determine a lower bound d_{low} on the value of d_{min} that is greater than zero, this bound must therefore be probabilistic as well. To state that Equation (2) should be true with large probability, define a small ϵ and set $\Pr[\forall i, j D_{i,j} >= d] = 1 - \epsilon$. Then we can develop Equation (2) as follows:

$$1 - \epsilon \approx (1 - \Pr[D <= d_{low} + 1])^{\frac{M^2-M}{2}}$$
$$(1 - \epsilon)^{\frac{2}{M^2-M}} \approx 1 - \Pr[D <= d_{low} + 1]$$

$$1 - (1 - \epsilon)^{\frac{2}{M^2 - M}} \quad \approx \quad \Pr[D <= d_{low} + 1].$$

Hence we conclude:

$$d_{low} \approx B^{-1} \left(1 - (1 - \epsilon)^{\frac{2}{M^2 - M}}, N, p_{neq} \right) - 1. \quad (3)$$

In the latter, the function $B^{-1}(q, n, p)$ is the inverse of the CDF of a binomial distribution with parameters n and p[1].

To construct an approximative upper bound d_{upp}, we state that with high probability at least one distance between any two codewords is smaller than a given d_{upp}:

$$1 - \epsilon \approx 1 - (1 - \Pr[D <= d_{upp} + 1])^{\frac{M^2 - M}{2}}. \quad (4)$$

Further development yields:

$$d_{upp} \approx B^{-1} \left(1 - \epsilon^{\frac{2}{M^2 - M}}, N, p_{neq} \right) - 1. \quad (5)$$

A mean value can be derived by setting the probability that all distances are greater than or equal to d_{avg} equal to the probability that there is a smaller distance:

$$\Pr[\forall i, j D_{i,j} >= d_{avg}] = \Pr[\exists i, j D_{i,j} < d_{avg}]. \quad (6)$$

The final formula takes the following form:

$$d_{avg} \approx B^{-1} \left(\frac{1}{2}, N, p_{neq} \right). \quad (7)$$

A measure is needed to express the level of individual classification errors that might occur when classifying members of the same class. We refer to the level of bit-errors between the binary representation class and the output of the classifiers for a given sample as the Signal-to-Noise ratio (SNR), analogous to the communication approach. The SNR is defined in terms of a Gaussian setup [1], i.e. Gaussian distributed codewords and Additive White Gaussian Noise (AWGN), and expressed in dB, i.e. SNR $= 10 \log_{10} \frac{\sigma_x^2}{\sigma_z^2}$ where σ_x^2 is the variance of the channel input and σ_z^2 is the variance of the noise. Under these conditions, the squared correlation coefficient between the channel input X and the channel output Y is $\rho_{X,Y}^2 = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_z^2}$. The relation between the SNR in a Gaussian setup and the probability of error per bit is then [7]:

$$\begin{aligned} p_b &= \frac{1}{\pi} \arccos \rho_{X,Y} \\ &= \frac{1}{\pi} \arccos \left(\sqrt{\frac{1}{1 + 10^{-\frac{\text{SNR}}{10}}}} \right). \quad (8) \end{aligned}$$

----
[1]Akin to the Matlab™function *binoinv(q, n, p).*

Conversely, a given p_b corresponds to an associated SNR by the following formula:

$$\text{SNR} = 10 \log_{10} \left(\frac{1}{\frac{1}{\cos^2(\pi p_b)} - 1} \right). \quad (9)$$

The definition of the SNR does not serve to show a link with a Gaussian setup per se, but rather to offer a frame of reference for the performance and efficiency of different setups.

3. ACHIEVABLE RATE WITH BOUNDED PROBABILITY OF ERROR

Let $T^{expected}$ be the number of erroneous bits that occur during classification with N binary classifiers with a probability of bit-error p_b, then $T^{expected} \sim Bin(N, p_b)$. Using a similar strategy as before, probabilistic bounds can be found on the number of erroneous bits. For a chosen small ϵ, lower and upper bounds can be established that are valid with large probability $1 - \epsilon$:

$$\begin{aligned} T_{low}^{expected} &= B^{-1}(\epsilon, N, p_b) \quad (10) \\ T_{upp}^{expected} &= B^{-1}(1 - \epsilon, N, p_b). \quad (11) \end{aligned}$$

Let the classification rate $R = \frac{\log_2(M)}{N}$ be a measure of efficiency akin to rates in communication. Essentially, the rate then shows what fraction of the used N bits is effectively needed to count the M different classes.

To guarantee that the number of errors does not exceed the number of errors tolerated, the condition $T_{upp}^{expected} \leq T_{low}^{tolerated}$ should be met. Calculating the corresponding rate for a given SNR requires that p_b is computed using Equation (8); then M can be computed for a chosen ϵ and a fixed N as:
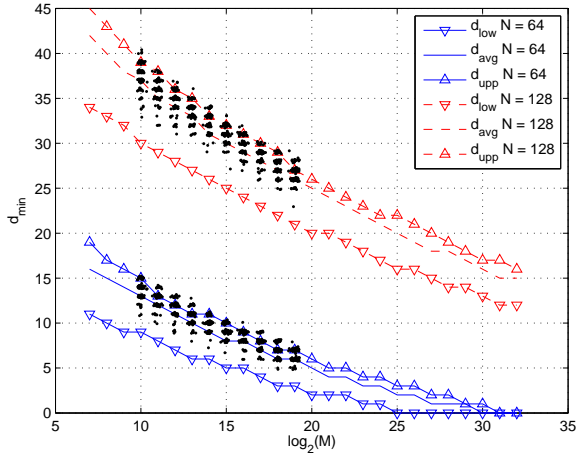
$$M = \frac{1 + \sqrt{1 + 4 \frac{2 ln(1 - \epsilon)}{ln \left(1 - B \left(2 B^{-1}(1 - \epsilon, N, p_b) + 2, N, p_{neq} \right) \right)}}}{2}. \quad (12)$$

The resulting probability of error can be bounded by requiring that for a success neither the number of bit-errors may exceed $T_{upp}^{expected}$, nor may d_{min} be smaller than d_{low}; both conditions are true with probability $1 - \epsilon$. Classification may still succeed when these conditions are not met but this is not guaranteed. By independence we then obtain:

$$\begin{aligned} p_e &= 1 - (1 - \epsilon)^2 \\ &\leq 2\epsilon. \quad (13) \end{aligned}$$

As a result, classification with a rate $R = \frac{\log_2(M)}{N}$ can be attained with a probability of error bounded by 2ϵ for a given SNR.

Fig. 1. Predictions and simulations 100 codebooks of 64- and 128-bit codewords with $p_1 = 0.5$ and $\epsilon = 0.001$ for different codebook sizes. Slight displacements have been added to the simulation results to show the distribution of the 100 values.
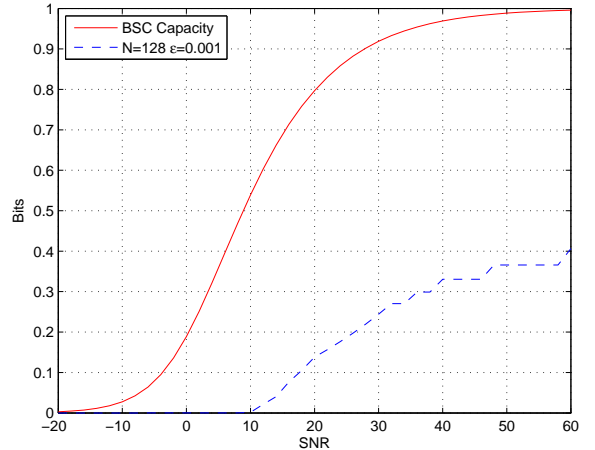
4. VALIDATION

4.1. The Maximum Number of Uniquely Distinguishable Classes

The accuracy of the estimation of the values of d_{min} has been verified for values of M that could be handled with the computers available to us. Unfortunately, the simulations have quadratic complexity and become prohibitively time-consuming for values of M that are still smaller than needed for a complete validation. This further stresses the need for an accurate analytical approach. In Figure 1 the predicted probabilistic bounds on d_{min} are depicted as lines for $N = 64$ and $N = 128$ as a function of M for $\epsilon = 0.001$ and $p_1 = 0.5$. The dots represent values for d_{min} found by simulation. The simulations were run 100 times for each value of M and frequently overlap, so to visualize the distribution of the values small random displacements have been added to the position of the dots.

The simulation results show that the proposed approximation is conservative, which is safe but may lead to slightly smaller estimated rates, or, equivalently, that for a given predicted rate the probability of error is smaller than the stated bounds.

4.2. Efficiency

The assumptions made in Section 2 about i.i.d. binary classifiers and and i.i.d. bit-errors imply full compatibility with a well-known setup in digital communications, namely the Binary Symmetric Channel (BSC) [1]. Theoretical results

Fig. 2. Predicted achievable rate for 128-bit codewords with $p_1 = 0.5$ and $\epsilon = 0.001$ versus the capacity of the Binary Symmetric Channel.

dictate that errorless communication can be accomplished with rates upto a certain maximum called the capacity C; under the assumption that the number of bits involved is asymptotically large and a minimum-Hamming distance decoder is used, $C = 1 + p_b \log_2 p_b + (1 - p_b) \log_2(1 - p_b)$. In the context of the present work, communication and rate are equivalent to classification and efficiency respectively, and the capacity presents an upper bound on the efficiency that can be attained.

There are however reasons why it is not realistic to expect that the capacity can truly be attained with real-life systems, for example that N is finite. As a result, the limit on the efficiency as predicted by our method should be lower.

A demonstration of the procedure described in Section 3 is shown in Figure 2, where an achievable rate for $N = 128$, $p_1 = 0.5$ and $\epsilon = 0.001$ is compared to the capacity of the BSC. There is a considerable gap between the achieved rate and the capacity, even for a probability of error bounded by 0.002 according to Equation (13), which is not very close to zero and should therefore be seen as a considerable relaxation with respect to the performance of a theoretical setup where the probability of error tends to zero. This confirms that systems based on a finite number of binary classifiers are not likely to reach the performance or efficiency one might expect based on classic communication-theoretic arguments.

Nonetheless, we must conclude that the current strategies leave room for considerable improvement: if M bits are used to distinguish between M classes, such as in the one-vs-all approach, the corresponding rate is but $\frac{\log_2 M}{M}$, which very rapidly approaches zero. For 128 classes, which is the situation depicted in Figure 2, a one-vs-all approach

has an efficiency of $\frac{7}{128} \approx 0.055$, and the all-vs-all approach scores $\frac{7}{8128} = 0.00086$, while the graph shows that greater efficiency is possible for SNRs greater than about 10 dB whilst maintaining a probability of classification error boudend by 0.002.

5. CONCLUSIONS

In Section 2 we have introduced a simple method to estimate the number of classes that can be distinguished using N binary classifiers for a chosen probability of classification error. In Section 3, the calculation of the efficiency or rate of a multi-class classifier has been shown. Finally, in Section 4, we have numerically confirmed the validity of the proposed estimation method and compared the efficiency of the one-vs-all and all-vs-all approaches to realistically attainable upper limits. The latter suggests that more efficient strategies must exist for most SNR-regimes greater than 10 dB.

6. REFERENCES

[1] Thomas Cover and Joy Thomas, *Elements of Information Theory*, John Wiley & Sons Inc., 2 edition, 2006.

[2] V.N.Vapnik, *Statistical Learning Theory*, John Wiley and Sons, 1998.

[3] Trevor Hastie and Robert Tibshirani, "Classification by pairwise coupling," *Annals of Statistics*, vol. 26, no. 2, pp. 451–471, 1998.

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

[5] Erin Allwein, Robert Schapire, and Yoram Singer, "Reducing multiclass to binary: A unifying approach for margin classifyers," *Journal of Machine Learning Research*, vol. 1, pp. 113–141, 2000.

[6] Ryan Rifkin and Aldebaro Klautau, "In defense of one-vs-all classification," *Journal of Machine Learning Research*, vol. 5, pp. 101–141, 2004.

[7] Sviatoslav Voloshynovskiy, Oleksiy Koval, Fokko Beekhof, and Thierry Pun, "Conception and limits of robust perceptual hashing: toward side information assisted hash functions," in *Proceedings of SPIE Photonics West, Electronic Imaging / Media Forensics and Security XI*, San Jose, USA, 2009.