EMPIRICALLY-ESTIMABLE MULTI-CLASS CLASSIFICATION BOUNDS

Alan Wisler[1], *Visar Berisha*[1], *Dennis Wei*[2], *Karthikeyan Ramamurthy*[2], *Andreas Spanias*[1,3]

[1]Arizona State University ECEE and SHS, [2]IBM Thomas J. Watson Research Center, [3]SenSIP Center

ABSTRACT

In this paper, we extend previously developed non-parametric bounds on the Bayes risk in binary classification problems to multi-class problems. In comparison with the well-known Bhattacharyya bound which is typically calculated by employing parametric assumptions, the bounds proposed in this paper are directly estimable from data, provably tighter, and more robust to different types of data. We verify the tightness and validity of this bound using an illustrative synthetic example, and further demonstrate its value by incorporating it into a feature selection algorithm which we apply to the real-world problem of distinguishing between different neuro-motor disorders based on sentence-level speech data.

Index Terms— Bayes error rate, multi-class classification, divergence measures, non-parametric estimator

1. INTRODUCTION

Supervised classification problems are based on the task of forming an approximate definition of an unknown labeling function $\phi(x)$ from a given sample of training data of the form $\{x_i, \phi(x_i)\}$ [1]. Problems in which $\phi(x) \in \{0, 1\}$ are referred to as binary classification problems and problems in which $\phi(x) \in \{1, 2, ...k\}$ are referred to as multi-class or k-class problems. Because binary classification problems are easier to solve, they have formed the test-bed for the development of most machine learning algorithms, while most multi-class approaches are extensions or generalizations of these binary solutions. In this manner a number of machine learning algorithms such as support vector machines [2][3], Neural Networks [4], and decision trees [5] have been generalized to multi-class problems.

An important aspect in the design of any predictive system, is the evaluation of its performance. Often, rather than comparing against a set of alternative classifiers, it would be preferable to compare against the optimal error rate. In the Bayesian setting when there exist known prior probabilities for each class, this can be represented by the Bayes Error Rate (BER) or Bayes Risk. If we consider two class distributions $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ in domain $\mathbf{x} \in \mathbb{R}^d$ with prior probabilities $p \in [0, 1]$ and $q = 1 - p$ respectively, the Bayes Risk can be defined as

$$P_{e01} = \int_{\mathbf{x}} \min\{pf_0(\mathbf{x}), qf_1(\mathbf{x})\}d\mathbf{x}. \quad (1)$$

This represents the error achieved by a classifier assigning a vector $\mathbf{x}$ the class with the highest posterior probability, and is the minimum error rate that can be achieved by any classifier.

The primary focus of this paper is to propose approaches for empirically estimating bounds on Bayes error in multi-class problems. In particular, we seek to extend non-parametric bounds on the Bayes error that can be calculated using the D_p divergence measure [6]. In order to achieve this, we utilize the principles behind the frameworks proposed in [7, 8] and apply them to the D_p divergence bound introduced in [9]. In a simulation based on synthetic data we compare these bounds to parametric and non-parametric estimates of multi-class bounds based on the Bhattacharyya coefficient (BC).We further examine the efficacy of these bounds in comparison to parametric bounds based on the Bhattacharyya distance by incorporating each into a feature selection (FS) algorithm. Previous work has shown that multi-class error bounds can be highly effective tool for dimensionality reduction [10, 11]. We apply each algorithm to the problem of discriminating between different neuro-motor disorders based on sentence-level speech data and compare the performance of classifiers constructed on each resulting subset of features.

2. BINARY CLASSIFICATION BOUNDS BASED ON THE D_P DIVERGENCE

Consider two class distributions $f_0(x)$ and $f_1(x)$ with prior probabilities $p \in [0, 1]$ and $q = 1 - p$ respectively. The D_p-divergence can be expressed as

$$D_p(f_0, f_1) = \frac{1}{4pq}\left[\int \frac{(pf_0(\mathbf{x}) - qf_1(\mathbf{x}))^2}{pf_0(\mathbf{x}) + qf_1(\mathbf{x})}d\mathbf{x} - (p-q)^2\right]. \quad (2)$$

This divergence measure was first introduced in [6], and can be directly estimated without estimation of $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ using an extension of the Friedman-Rafsky multivariate two sample test statistic [12]. The D_p-divergence has been used to estimate the Fisher information [6] as well as bounds on the BER and domain adaptation error for binary classification problems [9, 13]. In [9] it was shown that the BER for class distributions $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ can be bounded by

$$\frac{1}{2} - \frac{1}{2}\sqrt{u_p(f_0, f_1)} \leq P_{e01} \leq \frac{1}{2} - \frac{1}{2}u_p(f_0, f_1), \quad (3)$$

where

$$\begin{aligned} u_p(f_0, f_1) &= \int \frac{(pf_0(\mathbf{x}) - qf_1(\mathbf{x}))^2}{pf_0(\mathbf{x}) + qf_1(\mathbf{x})}d\mathbf{x} \\ &= 4pqD_p(f_0, f_1) + (p-q)^2. \end{aligned} \quad (4)$$

Here u_p is introduced for convenience, and is equivalent to D_p when $p = q = \frac{1}{2}$. These bounds have the nice properties of being non-parametric, empirically estimable, and provably tighter than the commonly used Bhattacharyya bounds [9].

3. EXTENDING BOUNDS TO MULTI-CLASS PROBLEMS

In this section we describe two multi-class extensions for the bounds introduced in Section 2: the first is a closed-form extension motivated by [7] and the second is a recursive multi-class extension based on the work in [8].

3.1. Closed-Form Extension

Consider an M-class problem with prior probabilities $p_1, ..., p_M$ and conditional class distributions $f_1(\mathbf{x}), ..., f_M(\mathbf{x})$ in hypothesis space

This research was supported in part by the Office of Naval Research grant N000141410722 (Berisha)

$\mathbf{x}$. We first consider extending the bounds using the approach described in [7]. In this paper, the authors show that the BER in multi-class ($\mathcal{R}^M$) problems can be bounded by

$$\frac{2}{M}\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}(p_i+p_j)P_{eij} \leq \mathcal{R}^M \leq \sum_{i=1}^{M-1}\sum_{j=i+1}^{M}(p_i+p_j)P_{eij} \tag{5}$$

where P_{eij} represents the pairwise Bayes risk of the 2-class subproblem of classifying between classes i and j. Substituting in the upper and lower bounds on the Bayes Risk defined in [9] yields

$$\begin{aligned}&\frac{2}{M}\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}(p_i+p_j)\left[\frac{1}{2}-\frac{1}{2}\sqrt{u_{\tilde{p}_i^{i,j}}(f_i(\mathbf{x}),f_j(\mathbf{x}))}\right]\\&\leq \mathcal{R}^M \leq \sum_{i=1}^{M-1}\sum_{j=i+1}^{M}(p_i+p_j)\left[\frac{1}{2}-\frac{1}{2}u_{\tilde{p}_i^{i,j}}(f_i(\mathbf{x}),f_j(\mathbf{x}))\right]\end{aligned} \tag{6}$$

where $\tilde{p}_i^{i,j}$ represents the normalized prior probability for class i defined by

$$\tilde{p}_i^{i,j} = \frac{p_i}{p_i+p_j}. \tag{7}$$

One limitation of this approach is that the upper bound becomes very loose when the overlap between class distributions is large. In fact, for completely overlapping distributions, the upper bound will converge to $(M-1)/2$ while the true BER converges to $(M-1)/M$. Section 3.2 will introduce an alternative that remedies this shortcoming.

3.2. Recursive Extension

Next we consider an expression introduced by Garber and Djouadi that represent bounds on the Bayes risk in terms of the Bayes risk of the M $(M-1)$-class subproblems created by removing different classes as

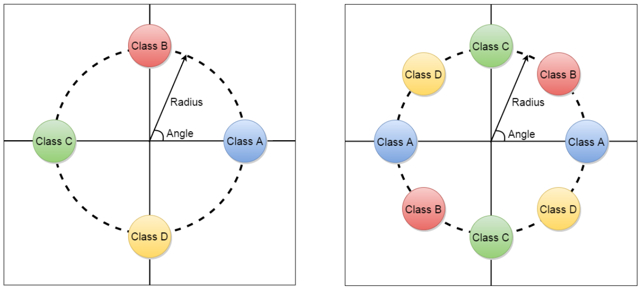
$$\begin{aligned}\frac{M-1}{(M-2)M}\sum_{i=1}^{M}(1-p_i)\mathcal{R}_i^{M-1} \leq \mathcal{R}^M \leq\\ \min_{\alpha\in\{0,1\}}\frac{1}{M-2\alpha}\sum_{i=1}^{M}(1-p_i)\mathcal{R}_i^{M-1}+\frac{1-\alpha}{M-2\alpha}\end{aligned} \tag{8}$$

Here $\mathcal{R}_i^{M-1}$ represents the Bayes risk for the $(M-1)$-class subproblem created by removing class i and α is an optimization constant that is minimized on-line in order to form the tightest possible bound. By using these upper and lower bounds in a recursive manner we can attain upper and lower bounds for the multi-class BER in terms of the pairwise Bayes risks between conditional class distributions. As in the first extension, we can bound each pairwise BER in terms of the D_p-divergence using (3). For example let us consider the 3-class case, we can compute the upper bound as

$$\mathcal{R}^3 \leq \min_{\alpha\in\{0,1\}}\frac{1}{3-2\alpha}\sum_{i=1}^{3}(1-p_i)\mathcal{R}_i^2+\frac{1-\alpha}{3-2\alpha}. \tag{9}$$

Substituting in the bounds expressed in (3) yields

$$\begin{aligned}\mathcal{R}^3 \leq \min_{\alpha\in\{0,1\}}\frac{1}{3-2\alpha}\Big\{&(p_1+p_2)\Big[\frac{1}{2}-\frac{1}{2}u_{\tilde{p}_1^{1,2}}(f_1,f_2)\Big]\\&+(p_1+p_3)\Big[\frac{1}{2}-\frac{1}{2}u_{\tilde{p}_1^{1,3}}(f_1,f_3)\Big]\\&+(p_2+p_3)\Big[\frac{1}{2}-\frac{1}{2}u_{\tilde{p}_2^{2,3}}(f_2,f_3)\Big]\Big\}+\frac{1-\alpha}{3-2\alpha}\end{aligned} \tag{10}$$



(a) Unimodal Scenario

(b) Bimodal Scenario

Fig. 1: Illustration of distribution placement for generating the synthetic data.

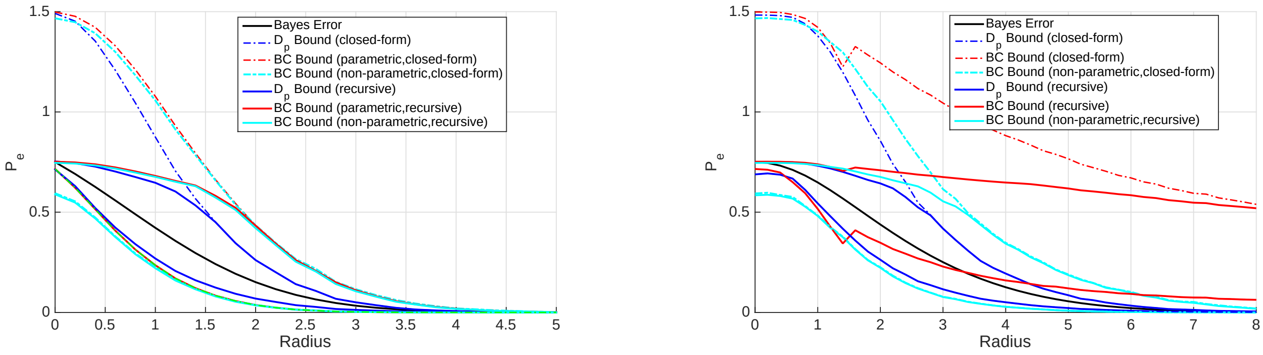
To better understand the role that α plays in this calculation, let us consider the two extreme cases in which the three class distributions are either completely overlapping or completely separable, and all class distributions have equal priors $p_1 = p_2 = p_3 = \frac{1}{3}$. In the first case, $\mathcal{R}_1^2 = \mathcal{R}_2^2 = \mathcal{R}_3^2 = \frac{1}{2}$ and $\alpha = 0$ yields the tightest bound of $\mathcal{R}^3 \leq \frac{2}{3}$ while $\alpha = 1$ yields the loosest bound of $\mathcal{R}^3 \leq 1$. In the second case $\mathcal{R}_1^2 = \mathcal{R}_2^2 = \mathcal{R}_3^2 = 0$, $\alpha = 0$ yields the loosest bound of $\mathcal{R}^3 \leq \frac{1}{3}$ while $\alpha = 1$ yields the tightest bound of $\mathcal{R}^3 \leq 0$. In general the value of α will depend on the total of the summation in (8). When this summation is greater than $(M-2)/2$ then $\alpha = 0$, otherwise $\alpha = 1$.

3.3. Comparison of Bounds

Because the two bounds are equivalent when $\alpha = 1$, Garber was able to show that for problems with equal priors the recursive extension is guaranteed to be at least as tight as the closed-form extension [8]. Extended proofs in Section 6 shows both the upper and lower recursive bounds will be at least as tight as the closed-form bounds regardless of priors. The price for this superiority comes in the increased computational burden. The computational burden of the closed-form bound can be approximated by $M(M-1)\gamma(n_c)/2$, where $\gamma(n_c)$ represents the number of computations required for a single pairwise risk function between classes containing n_c samples. In addition to these computations, the recursive bound requires calculation of (8) for all $\sum_{i=3}^{M-1}\binom{M}{i}$ unique subproblems of 3 or more classes. While these additional computations are inconsequential for small M, their rapid growth w.r.t. M makes this method infeasible for problems containing a large number of classes ($M > 30$).

4. RESULTS

This section is divided into two parts. Section 4.1 presents an illustrative example with synthetic data in which several class distributions are represented by either unimodal or bimodal Gaussian distributions. The unimodal case illustrates the tightness of bounds based on the D_p divergence relative to bounds based on the Bhattacharyya distance, as well as the differences in the two extension methods discussed in Section 3. The bimodal case illustrates the vulnerability of parametric bounds to non-Gaussian data. In Section 4.2 we utilize a feature selection algorithm based on these bounds to identify feature subsets that discriminate between different speech disorders.



(a) Unimodal Scenario

(b) Bimodal Scenario

Fig. 2: True BER and error bounds for varying radii generated by each scenario of the synthetic data simulation.

4.1. Synthetic Data Example

To test the accuracy of the proposed bounds we consider the scenario in which four bivariate class distributions are equally spaced in a radial formation around the origin. We consider two scenarios. In the first scenario, each class distribution is represented by a single Gaussian distribution. In the second scenario the class distributions from the first scenario are augmented by a second Gaussian distribution at a 180° rotation from the first. This second scenario is used to illustrate the behavior of the Bhattacharyya bound when the parametric assumption that each class can be modeled by a single Gaussian does not fit the actual data. The distribution placements used in each scenario are presented in Figure 1.

Throughout this experiment, each Gaussian is isotropic with unit covariance, and mean determined by the angle and radius. The angle used to place each distribution is held constant (see Figure 1) while the radius is varied from zero, where the distributions in each scenario are completely overlapping, to eight where the distributions in each scenario contain almost no overlap with the neighboring distributions. The radius is varied in increments of 0.2, and each class distribution is represented by 1000 samples of data generated according to the parameters of the distribution. At each radius, we generate bounds on the Bayes error using both the recursive and closed-form extensions described in Section 3 for the D_p and BC bounds. In these calculations, the D_p-divergences are calculated using the approach described in [9]. The Bhattacharyya distances are estimated in a parametric fashion by empirically estimating the mean and covariance matrices, then plugging the results into the explicit formula for multivariate normal distributions defined in [14], and in a non-parametric fashion by using a 2-dimensional histogram to estimate each underlying distribution and solving for the BC by integration. We obtain a ground truth value of the BER by integrating across the true underlying class distributions. To reduce the variance of the estimator we average our results across 25 Monte Carlo iterations. The resulting bounds are shown in Figure 2.

In Figure 2a, we see little difference between the parametric and non-parametric estimates of the Bhattacharyya bound, other than a slight negative bias that is most pronounced for tightly overlapping distributions. Figure 2b shows that while the non-parametric D_p and BC bounds remain largely unaffected by the addition of the second Gaussian for each class, the parametric bounds do not hold for radii exceeding 1.5 when the separation between modes is sufficient to violate the parametric assumption. In both scenarios, the D_p bound provides a tighter bound on the BER. It should be noted that the benefits of the D_p bound will only become more pronounced in high-dimensional spaces where accurate non-parametric density estimation is often infeasible [15].

4.2. Disordered Speech Example

Dysarthria is a motor speech disorder resulting from an underlying neurological injury. In this Section, we discuss the challenge of distinguishing between three different Dysarthrias: Parkinson's, Amyotrophic Lateral Sclerosis (ALS), and Ataxic Dysarthria. Automatically classifying between different neurogenic disorders from speech presents a major engineering challenge.

4.2.1. Data

We make use of data collected in the Motor Speech Disorders Laboratory at Arizona State University, consisting of 71 dysarthric speakers. Among these speakers we examine 17 speakers with ataxic dysarthria, secondary to cerebellar degeneration, 15 mixed flaccid-spastic dysarthria, secondary to ALS, and 39 speakers with hypokinetic dysarthria secondary to Parkinson's Disease. Each patient provided speech samples, including a reading passage, phrases, and sentences. The speech database consists of approximately 10 minutes of recorded material per speaker. For a more detailed description of the methods used to collect this dataset see [16].

4.2.2. Experiment

We partitioned the database into training and test sets, by randomly selecting 10 speakers from each subtype and 20 sentences from each speaker to be placed in the training set. Complete sentence data from all remaining speakers is then assigned to the test set. After partitioning the data, we extract a total of 1201 features including 99 long-term average spectrum (LTAS) features [17], 60 Envelope Modulation Spectrum (EMS) features [18], 234 mel frequency cepstral coefficients (MFCC) features, and 783 additional spatio-temporal features [19]. We then iteratively select features using a forward selection feature selection (FS) algorithm that attempts to minimize the closed-form and recursive extensions of the parametric Bhattacharyya and non-parametric D_p bounds discussed in Section 4.1. We also include a wrapper feature selection method that iteratively selects the features that maximize the performance of the classifier on a held-out validation set. Wrappers will typically identify the optimal subset of features for the selected classifier, but are

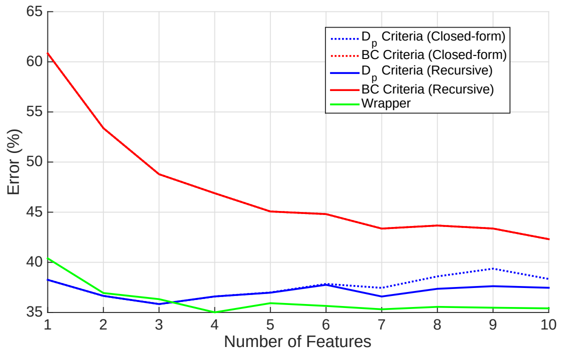


Fig. 3: Error rates in distinguishing between different speech disorders as a function of the number of features selected.

computationally very burdensome [20] (run time exceeds 5 times that of the proposed algorithm). Each FS algorithm is used to identify feature subsets of sizes 1-10. For each subset a classification tree is trained on the training data, and evaluated on the test data. This entire procedure is repeated over a 20-iteration Monte Carlo simulation and the average performance achieved by the subsets from each FS algorithm is displayed in Figure 3.

Figure 3 shows that the D_p-based FS algorithm achieved superior performance to BC-based algorithm throughout the experiment, although the gap narrows as additional features are added. While the D_p algorithm achieves slightly higher performance in the smaller subsets, the wrapper yields the highest overall performance. We were not able to observe any significant difference in the closed-form and recursive bounds in this experiment, and other than some of the later features chosen by the D_p algorithm the two methods generally returned the same set of features. This indicates that we are operating in the regime in Figure 2 after the two methods have converged and the bounds become virtually identical.

5. CONCLUSION

In this paper we examine two previously established methods of bounding the Bayes risk in multi-class machine learning problems. Using these methods we generalize binary classification bound based on the D_p-divergence to multi-class problems. We demonstrate the tightness of this bound in the multi-class setting in an experiment using synthetic data. We then examine the efficacy of a feature selection algorithm based on this bound for the classification of different neuro-motor disorders based on sentence-level speech data.

6. APPENDIX

6.1. Upper bound proof

To prove that the recursive bound is tighter than the closed-form bound, it is sufficient to prove that $\Phi^M = \Theta^M \quad \forall M$, where Φ^M represents the recursive bound when $\alpha = 1$, and Θ^M represents the closed-form upper bound.

Basic Step: Prove that $\Phi^3 = \Theta^3$

$$\Phi^3 = \frac{1}{3-2\alpha}\sum_{i=1}^{3}(1-p_i)\Phi_i^2 = \Theta^3 \tag{11}$$

Inductive Step: Suppose that $\Phi^{M-1} = \Theta^{M-1}$. By definition:

$$\Phi^M = \frac{1}{M-2}\sum_{i=1}^{M}(1-p_i)\Phi_i^{M-1} \tag{12}$$

Using the inductive hypothesis

$$\begin{aligned}\Phi^M &= \frac{1}{M-2}\sum_{i=1}^{M}(1-p_i)\Theta_i^{M-1}\\ &= \frac{1}{M-2}\sum_{i=1}^{M}(1-p_i)\sum_{\substack{j=1\\ j\neq i}}^{M}\sum_{\substack{k=j+1\\ k\neq i}}^{M}\left(\frac{p_j}{1-p_i}+\frac{p_k}{1-p_i}\right)P_{ejk}\end{aligned} \tag{13}$$

where $\frac{p_j}{1-p_i}$ reflects the normalized prior probability of class j for the $(M-1)$-class subproblem with class i removed. After canceling the $(1-p_i)$ terms

$$\Phi^M = \frac{1}{M-2}\sum_{i=1}^{M}\sum_{\substack{j=1\\ j\neq i}}^{M}\sum_{\substack{k=j+1\\ k\neq i}}^{M}(p_j+p_k)P_{ejk} \tag{14}$$

Note that every pairwise Bayes risk P_{ejk} will occur in the interior double summation except for those containing $j = i$ or $k = i$. Therefor every pairwise Bayes risk $P_{ejk}, j \in [M], k \in [M\backslash j]$ will occur in the triple summation $(M-2)$-times (M-times minus the two instances when $j = i$ or $k = i$). We can thus remove the outer summation and the expression simplifies to

$$\begin{aligned}\Phi^M &= \frac{1}{M-2}\sum_{j=1}^{M}\sum_{k=j+1}^{M}(M-2)(p_j+p_k)P_{ejk}\\ &= \sum_{j=1}^{M}\sum_{k=j+1}^{M}(p_j+p_k)P_{ejk} = \Theta^M.\end{aligned} \tag{15}$$

Therefore, by induction $\Phi^M = \Theta^M \quad \forall M$, and the recursive bound must be at least as tight as the closed-form bound.

6.2. Lower bound proof

Prove that the recursive lower bound (ϕ^M) equals the closed-form lower bound(θ^M)

$$\theta^M = \frac{2}{M}\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}(p_i+p_j)P_{eij} \tag{16}$$

Basic Step:

$$\phi^3 = \frac{2}{3}\sum_{i=1}^{3}(1-p_i)\phi_i^2 = \frac{2}{3}\sum_{i=1}^{2}\sum_{j=i+1}^{3}(p_i+p_j)P_{eij} \tag{17}$$

Inductive Step: Suppose $\phi^{M-1} = \theta^{M-1}$, substituting this into the definition for ϕ^M yields:

$$\begin{aligned}\phi^M &= \frac{M-1}{M(M-2)}\sum_{i=1}^{M}\frac{2(1-p_i)}{M-1}\sum_{\substack{j=1\\ j,k\neq i}}^{M-1}\sum_{k=j+1}^{M}\frac{(p_j+p_k)}{1-p_i}P_{ejk}\\ &= \frac{2}{M(M-2)}\sum_{i=1}^{M}\sum_{\substack{j=1\\ j,k\neq i}}^{M-1}\sum_{k=j+1}^{M}(p_j+p_k)P_{ejk}\\ &= \frac{2}{M}\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}(p_i+p_j)P_{eij}\end{aligned} \tag{18}$$

Therefore, by induction the recursive and closed-form lower bounds are equivalent for all M.

7. REFERENCES

[1] Thomas G. Dietterich and Ghulum Bakiri, “Solving multiclass learning problems via error-correcting output codes,” *Journal of artificial intelligence research*, pp. 263–286, 1995.

[2] Koby Crammer and Yoram Singer, “On the algorithmic implementation of multiclass kernel-based vector machines,” *The Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2002.

[3] Erin J Bredensteiner and Kristin P Bennett, “Multicategory classification by support vector machines,” in *Computational Optimization*, pp. 53–79. Springer, 1999.

[4] Christopher M Bishop, *Neural networks for pattern recognition*, Oxford university press, 1995.

[5] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen, *Classification and regression trees*, CR-C press, 1984.

[6] Visar Berisha and Alfred O Hero, “Empirical non-parametric estimation of the Fisher information,” *Signal Processing Letters, IEEE*, vol. 22, no. 7, pp. 988–992, 2015.

[7] Tsvi Lissack and King-Sun Fu, “Error estimation in pattern recognition via L_α-distance between posterior density functions,” *Information Theory, IEEE Transactions on*, vol. 22, no. 1, pp. 34–45, 1976.

[8] FD Garber and Abdelhamid Djouadi, “Bounds on the Bayes classification error based on pairwise risk functions,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 10, no. 2, pp. 281–288, 1988.

[9] Visar Berisha, Alan Wisler, Alfred O Hero, and Andreas S-panias, “Empirically estimable classification bounds based on a new divergence measure,” *Signal Processing, IEEE Transactions on*, 2015, In press.

[10] Luis Rueda and Myriam Herrera, “A new approach to multiclass linear dimensionality reduction,” in *Progress in Pattern Recognition, Image Analysis and Applications*, pp. 634–643. Springer, 2006.

[11] Madan Thangavelu and Raviv Raich, “Multiclass linear dimension reduction via a generalized chernoff bound,” in *Machine Learning for Signal Processing, 2008. MLSP 2008. IEEE Workshop on*. IEEE, 2008, pp. 350–355.

[12] Jerome H Friedman and Lawrence C Rafsky, “Multivariate generalizations of the Wald-Wolfowitz and smirnov two-sample tests,” *The Annals of Statistics*, pp. 697–717, 1979.

[13] Alan Wisler, Visar Berisha, Julie Liss, and Andreas Spanias, “Domain invariant speech features using a new divergence measure,” in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 77–82.

[14] Thomas Kailath, “The divergence and Bhattacharyya distance measures in signal selection,” *Communication Technology, IEEE Transactions on*, vol. 15, no. 1, pp. 52–60, 1967.

[15] Alfred O Hero, Bing Ma, Olivier Michel, and John Gorman, “Alpha-divergence for classification, indexing and retrieval,” *Communication and Signal Processing Laboratory, Technical Report CSPL-328, U. Mich*, 2001.

[16] Kaitlin L Lansford and Julie M Liss, “Vowel acoustics in dysarthria: Speech disorder diagnosis and classification,” *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 1, pp. 57–67, 2014.

[17] Phil Rose, *Forensic speaker identification*, CRC Press, 2003.

[18] Julie M Liss, Sue LeGendre, and Andrew J Lotto, “Discriminating dysarthria type from envelope modulation spectra,” *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 5, pp. 1246–1255, 2010.

[19] James R Williamson, Thomas F Quatieri, Brian S Helfer, Rachelle Horwitz, Bea Yu, and Daryush D Mehta, “Vocal biomarkers of depression based on motor incoordination,” in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 41–48.

[20] Ron Kohavi and George H John, “Wrappers for feature subset selection,” *Artificial intelligence*, vol. 97, no. 1, pp. 273–324, 1997.