RATE-OPTIMAL META LEARNING OF CLASSIFICATION ERROR

Morteza Noshad and Alfred O. Hero III

University of Michigan, Electrical Engineering and Computer Science, Ann Arbor, Michigan, U.S.A

ABSTRACT

Meta learning of optimal classifier error rates allows an experimenter to empirically estimate the intrinsic ability of any estimator to discriminate between two populations, circumventing the difficult problem of estimating the optimal Bayes classifier. To this end we propose a weighted nearest neighbor (WNN) graph estimator for a tight bound on the Bayes classification error; the Henze-Penrose (HP) divergence. Similar to recently proposed HP estimators [1], the proposed estimator is non-parametric and does not require density estimation. However, unlike previous approaches the proposed estimator is rate-optimal, i.e., its mean squared estimation error (MSEE) decays to zero at the fastest possible rate of $O(1/M + 1/N)$ where M, N are the sample sizes of the respective populations. We illustrate the proposed WNN meta estimator for several simulated and real data sets.

1. INTRODUCTION

The minimum classification error, also known as Bayes error, is the best (minimum) average probability of error that can be achieved by any binary classifier of a sample coming from one of two classes. Although the Bayes error can be represented in the form of an integral for a simplest case of binary classification, analytical evaluation of this integral is often not feasible, even when densities are known [1]. Based on this fact, several previous works have investigated upper and lower bounds on the Bayes error, which are easy to compute analytically. A Bound based on Chernoff α-divergence has been proposed in [2]. Bhattacharya divergence, which is a special case of Chernoff α-divergence for $\alpha = \frac{1}{2}$, is used in a number of applications involving Bayes error including feature selection [3–5]. Recently Berisha et al proposed tighter lower and upper bounds based on HP divergence with parameter p, where p is the prior probability of class 1 [1]. They proved that the bounds are tight for $p = 1/2$.

The problem of estimating the Bayes classification error directly, without the need to estimate the Bayes classifier function, is called the meta-learning problem. It is of crucial importance in reinforcement learning where an estimate of potential performance gains is used to guide the choice of future actions, e.g., selecting a data source [6]. Information divergence approaches to solving the meta-learning problem estimate various divergence functionals that measure the dissimilarity between the population distributions.

There are two major information divergence estimation approaches; plug-in and direct estimation. Plug-in methods first compute estimates of the population densities and plug them into the formula for the information divergence. Kernel Density Estimator (KDE) and k-Nearest Neighbor (k-NN) population density estimation are commonly used in the plug-in approach [7, 8]. In contrast, direct estimation approaches bypass density estimation entirely, producing an estimator of the information divergence using geometric functions of the data. Geometric quantities such as minimal graphs are commonly used for direct estimation of information divergences using the k-NN graph on the dataset to estimate Rényi entropy [9,10], using the MST graphs to estimate HP divergence [11], and using the nearest neighbor ratios (NNR) method to estimate various divergence measures [12], are some of the examples of direct graph based estimators. Direct estimation methods have several advantages over plug-in estimators such as lower computational complexity, simplicity of the estimator, imposing less constraints on the density functions, and offering an intuitive graph theoretical interpretation of the information measure.

The HP divergence was defined by Henze [13, 14] as the almost sure limit of the Friedman-Rafsky (FR) multi-variate two sample test statistic. Thus the FR two sample test statistic can be interpreted as an asymptotically consistent estimator of the HP divergence. The FR procedure is as follows. Assume that we have two data-sets X and Y. The FR test statistic is formed by counting the edges of MST graph of the joint data set $Z := X \cup Y$, which connect dichotomous points, i.e., a point in X to a point in Y. Later in [14], Henze proposed another similar graph based estimator that considers k-NN graph instead of the MST graph. However, the main FR test statistics using MST graph has received more attention than the k-NN variant. The authors of [14] proved the asymptotic consistency of FR statistics based on type coincidence, but the convergence rates of these estimators have remained unknown since then. In [15] Moon et al used an optimal plug-in density estimator to estimate HP divergence. As mentioned before, since plug-in estimation needs multi-step estimation procedure, which consists of estimating each of the densities in the first place, and then plugging-in the estimated densities

This work was partially supported by a grant from ARO, number W911NF-15-1-0479.

in the divergence function, compared to the direct estimation methods, it suffers from slower runtimes and requires stringent on the density functions.

In this paper we propose a new direct estimator of the HP divergence based on a weighted k-NN graph. We first derive the convergence rates of the k-NN based FR test statistics, defined as the number of edges in the k-NN graph over the joint data set $Z := X \cup Y$, which connect dichotomous points. We prove that the bias rate of this estimator is upper bounded by $O\Big((k/N)^{\gamma/d}\Big) + e^{-ck}$, where N and d respectively are the number and dimension of the samples, γ is the Hölder smoothness parameter of the densities and c is a constant. Note that the convergence rate of this estimator worsens in higher dimensions and does not achieve the optimal parametric rate of $O(1/N)$. Therefore, we propose a direct estimation method based on a weighted k-NN graph. We refer to this method as the weighted nearest neighbor (WNN) estimator. The graph includes a weighted, directed edge between any pair of nodes R and S only if the types of R and S are different (i.e. $R \in X$ and $S \in Y$) and S belongs to the set of kth nearest neighbors of R. We prove that if the edge weights are obtained from the solution of a certain optimization problem, we can construct a rate-optimal HP divergence estimator based on the sum of the weights of the dichotomous edges. The convergence rate of this estimator is proved to be $O(1/N)$, which is both optimal and independent of d. Finally, we emphasize that the proposed WNN estimator is completely different from the weighted matching estimator.

2. MAIN RESULTS

In this section we recall the Henze-Penrose (HP) divergence and propose an optimal estimator based on the k-NN graph. All of the proofs of the convergence theorems are provided in Appendix of [16] (Section 6).

Consider two density functions f_X and f_Y with support $\mathcal{M} \subseteq \mathbb{R}^d$. The HP-divergence between f_X and f_Y is denoted by $D_P\left(f_X(x)||f_Y(x)\right)$ and defined as

$$D_p = 1 - \int \frac{f_X(x) f_Y(x)}{p f_X(x) + q f_Y(x)} dx \tag{1}$$

where p is a parameter and $p + q = 1$. We also define $\eta := p/q$. In [12] p is the number of empirical samples from the first class.

Assumptions: We assume that the densities f_1 and f_2 have the same bounded support set and are lower bounded by $C_L > 0$ and upper bounded by C_U. We also assume that they belong to Hölder smoothness class with parameter γ:

Definition Given the support set $\mathcal{X} \subseteq \mathbb{R}^d$, a function $f : \mathcal{X} \to \mathbb{R}$ is called Hölder continuous with parameter $0 < \gamma \leq 1$, if there exists a positive constant G_f, possibly depending on f, such that for every $x \neq y \in \mathcal{X}$,

$$|f(y) - f(x)| \leq G_f \|y - x\|^{\gamma}, \tag{2}$$

2.1. k-NN Estimator

Definition Let $X = \{X_1, ..., X_N\}$ and $Y = \{Y_1, ..., Y_M\}$ respectively denote i.i.d samples with densities f_1 and f_2, such that $M = \left\lfloor \frac{Nq}{p} \right\rfloor$. Let $G_k(X, Y)$ be the graph of k nearest neighbors constructed over the joint set $Z = X \cup Y$. In other words, in $G_k(X, Y)$, each point $x \in Z$ is connected to its kth nearest neighbors, denoted by $Q_k(x)$. Assume that $\mathcal{E}(X, Y)$ is the set of edges of $G_k(X, Y)$ connecting dichotomous points. Then the K-NN estimator of HP-divergence, $\widehat{D_p}$, is defined as

$$\widehat{D_p}(X, Y) = 1 - |\mathcal{E}(X, Y)| \frac{N + M}{2NM}. \tag{3}$$

The idea behind this estimator is similar to the idea of MST estimator of HP-divergence proposed by Friedman and Rafsky (FR) [17], in which we count the number of edges connecting dichotomous points in the minimal spanning tree of the joint data [13]. If $N = M$ and the densities are almost equal, then with probability of almost $1/2$ every kth nearest neighbor edge belongs to $\mathcal{E}$(X,Y). In this case $|\mathcal{E}(X, Y)| \approx N$, and $\widehat{D_p} \approx 0$. The algorithm for k-NN estimator is provided in Algorithm 1. By using the Kd-tree method for construction of the k-NN graph, the computational complexity of this algorithm is $O(kN \log N)$ [12].

In the following theorems we derive upper bounds on the bias and variance rates. Here the bias and variance are defined as $\mathbb{B}[\hat{T}] = \mathbb{E}[\hat{T}] - T$ and $\mathbb{V}[\hat{T}] = \mathbb{E}[\hat{T}^2] - \mathbb{E}[\hat{T}]^2$, respectively, where $\hat{T}$ is an estimator of the parameter T.

Theorem 2.1 *The bias of the k-NN estimator for HP divergence satisfies*

$$\mathbb{B}\Big[\widehat{D_p}(X, Y)\Big] = O\Big((k/N)^{\gamma/d}\Big) + O(\mathcal{C}(k)), \tag{4}$$

where $\mathcal{C}(k) := exp(-3k^{1-\delta})$ for a fixed $\delta \in (2/3, 1)$. Here γ is the Hölder smoothness parameter.

Remark 1 *Note that in order that $\hat{D}_p(X, Y)$ be asymptotically unbiased, k needs to grow with N. The minimum bias rate of $O\Big(\frac{\log N}{N}\Big)^{\gamma/d}$ can be achieved by selecting $k = O(\log N)$.*

Theorem 2.2 *The variance of the k-NN estimator for the HP divergence satisfies*

$$\mathbb{V}\Big[\widehat{D_p}(X, Y)\Big] \leq O\Big(\frac{1}{N}\Big). \tag{5}$$

2.2. WNN Estimator

Note that the bias term in Theorem 2.1 depends on d. Therefore, for higher dimensions the estimator convergence rate is slower. In order to resolve this issue and achieve optimum convergence rate in any dimension, we propose a modified k-NN graph based estimator of HP divergence. Assume

Algorithm 1: k-NN Estimator of HP Divergence

Input : Data sets $X = \{X_1, ..., X_N\}$, $Y = \{Y_1, ..., Y_M\}$

$Z \leftarrow X \cup Y$
for *each point* Z_i *in* Z **do**
  If ($Z_i \in X$ and $Q_k(Z_i) \in Y$)
    or ($Z_i \in Y$ and $Q_k(Z_i) \in X$)
      then $t \leftarrow t + 1$

Output: $1 - t\frac{N+M}{2NM}$

Algorithm 2: WNN Estimator of HP Divergence

Input : Data sets $X = \{X_1, ..., X_N\}$, $Y = \{Y_1, ..., Y_M\}$

$Z \leftarrow X \cup Y$
for $l \in \mathcal{L}$ **do**
  for *each point* Z_i *in* Z **do**
    If ($Z_i \in X$ and $Q_l(Z_i) \in Y$)
      or ($Z_i \in Y$ and $Q_l(Z_i) \in X$)
        then $t \leftarrow t + W(l)$

Output: $1 - t\frac{N+M}{2NM}$

that the density functions are in the Hölder space $\Sigma(\gamma, B)$, which consists of functions on $\mathcal{X}$ continuous derivatives up to order $q = \lfloor\gamma\rfloor \geq d$ and the qth partial derivatives are Hölder continuous with exponent $\gamma' =: \gamma - q$ and Lipshitz constant B. Further, assume that the density derivatives up to order d vanish at the boundary. Fix a constant L where $L \geq d$. Let $\mathcal{L} := \{l_1, ..., l_L\}$ be a set of index values with $l_i < c$, where c is a constant. For instance one can assume that $\mathcal{L} := \{1, ..., d\}$. We further define the value of the k-NN parameter as a function of l, i.e. $K(l) := \left\lfloor l\sqrt{N} \right\rfloor$.

Definition Let $X = \{X_1, ..., X_N\}$ and $Y = \{Y_1, ..., Y_M\}$ respectively denote i.i.d samples with densities f_1 and f_2, such that $M = \left\lfloor \frac{Nq}{p} \right\rfloor$. Let the weight vector $W := [W(l_1), W(l_2), ..., W(l_L)]$ be the solution to the following optimization problem:

$$\begin{aligned} \min_{w} \quad & \|w\|_2 \\ \textit{subject to} \quad & \sum_{l\in\mathcal{L}} w(l) = 1, \\ & \sum_{l\in\mathcal{L}} w(l) l^{i/d} = 0, i \in \mathbb{N}, i \leq d. \end{aligned} \tag{6}$$

Now define $G_K^W(X, Y)$ as a weighted directed graph over vertices of the joint set $X \cup Y$. There is a directed edge with the weight $W(l)$ between any pair of nodes R and S, only if the types of R and S are different (i.e. $R \in X$ and $S \in Y$), where S is the $K(l)$-th nearest neighbor of R for some $l \in \mathcal{L}$. We represent the set of edges of $G_K^W(X, Y)$ by $\mathcal{E}_K^W(X, Y)$.

The proposed WNN estimator $\widehat{D}_p^W$ of HP divergence, is defined as

$$\widehat{D}_p^W(X, Y) = 1 - |\mathcal{E}_K^W(X, Y)|\frac{N + M}{2NM}. \tag{7}$$

Note that the weights in (6) are computed offline as the solution of a linear programming problem. The algorithm for the WNN estimator is provided in Algorithm 2.

Theorem 2.3 *The MSE of the WNN estimator is* $O(1/N)$.

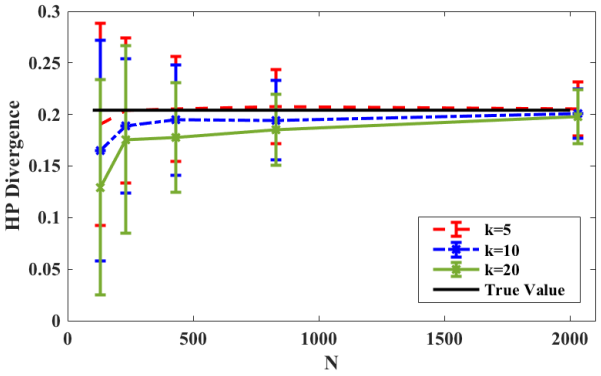

Fig. 1. Comparison of the estimated values of k-NN estimator with $k = 5, 10, 20$ for HP divergence between two truncated Normal RVs with mean vectors $[0, 0]$ and $[0, 1]$ and variances of $\sigma_1^2 = \sigma_2^2 = I_2$, plotted against N, the number of samples.

3. NUMERICAL RESULTS

In this section we investigate the behavior of the proposed estimator by numerical experiments.

Fig. 1 shows the mean estimated HP divergence between two truncated Normal RVs with mean vectors $[0, 0]$ and $[0, 1]$ and variance of $\sigma_1^2 = \sigma_2^2 = I_2$, as a function of number of samples, N, where I_d is the identity matrix of size d. Three different values of k are investigated. For each case we repeat the experiment 100 times, and compute the bias and variance. As N increases, the bias for any k tends to zero. The experiments show that the bias decreases slower as k increases, which is due to the $O\left(\left(\frac{k}{N}\right)^{\gamma/d}\right)$ term in (4). However, according to this experiment, the variance is almost independent of k and decreases linear towards zero.

Fig. 2 shows the MSE of the k-NN estimator for HP divergence between two zero mean Normal random vectors in $\mathbb{R}^2$, with identical covariance matrix I_d whose values are truncated within the range $x \in [-5, 5]$ and $y \in [-5, 5]$. The experiment is repeated for three different dimensions of $d = 2, 10, 20$ for fixed $k = 5$. In agreement with our bias bound in (4), as d increases, the experiment shows that the

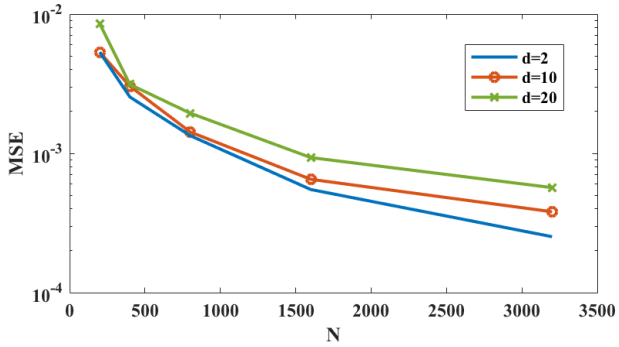

Fig. 2. MSE of the k-NN estimator for HP divergence between two identical, independent and truncated Normal RVs, as a function of N.

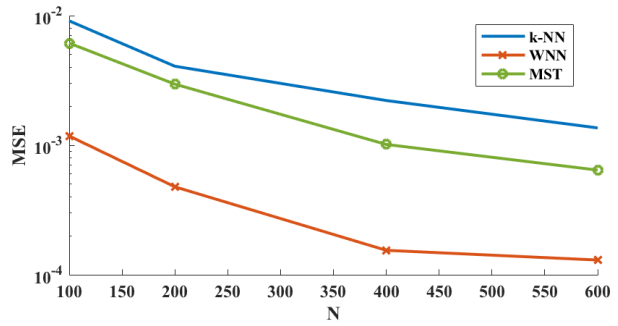

Fig. 3. MSE comparison of the three graph theoretical estimators of HP divergence; MST, k-NN, and WNN estimators.

MSE rate increases.

In Fig. 3 we compare the MSE rates of the three graph theoretical estimators of HP divergence; MST, k-NN, and WNN estimators. The HP divergence between two truncated Normal random variables with $d = 2$, means of $\mu_1 = [0, 0]$, $\mu_2 = [1, 0]$, and covariance matrices of $\sigma_1 = I_2$ and $\sigma_2 = 2I_2$. This experiment verifies the advantage of WNN estimator over the k-NN and MST estimators, in terms of their convergence rates. Also the performance of MST estimator is slightly better than the k-NN estimator. Note that in this experiment we have set the number of neighbors of the k-NN to $k = 5$.

Fig. 4 shows the comparison of the estimators of HP divergence between a truncated Normal RV with mean $[0, 0]$ and covariance matrix of I_2, and uniform RV within $[-5, 5] \times [-5, 5]$, in terms of their mean value and $\%95$ confidence band. The confidence band becomes narrower for greater values of N, and the WNN estimator has the narrowest confidence band.

Finally in Fig. 5, we compare performance of the WNN to that of the k-NN estimators with $k = 5$ and $k = 10$, for a real data set [18, 19]. The data are measurement from a set of ultrasound sensors arranged circularly around a robot,

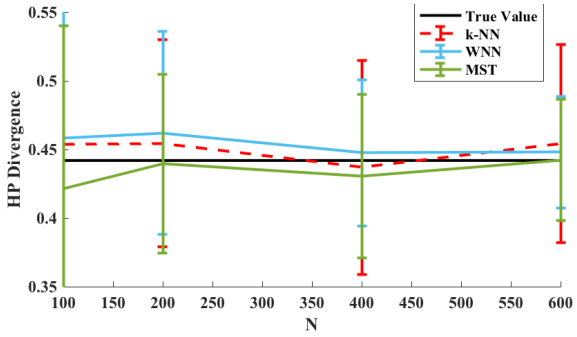

Fig. 4. Comparison k-NN, MST and WNN estimators of HP divergence between a truncated Normal RV and a uniform RV, in terms of their mean value and $\%95$ confidence band.

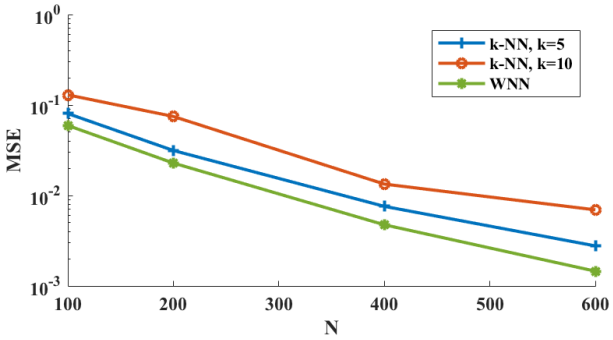

Fig. 5. MSE Comparison of the WNN and k-NN estimator with two different parameters $k = 5$ and $k = 10$ for the robot navigation dataset.

which navigates through the room following the wall in a clockwise direction. There are total number of 5456 instances (corresponding to different timestamps), and we use the information of four main sensors as the feature space. The instances are associated with four different classes of actions: move-forward, sharp-right-turn, slight-right-turn and turn-left. In Fig. 5 we consider the divergence between the sensor measurement for sharp-right-turn and move-forward classes. Note the superior performance of the WNN estimator as compared to the k-NN estimators.

4. CONCLUSION

In this paper we derived the convergence rates of k-NN version of FR test statistics and proposed an optimum direct estimation method for HP divergence, based on the weighted k-NN graph. We proved that WNN estimator can achieve optimum parametric MSE rate of $O(1/N)$, and we validated our results on simulated and real data sets.

5. REFERENCES

[1] Visar Berisha, Alan Wisler, Alfred O Hero, and Andreas Spanias, "Empirically estimable classification bounds based on a nonparametric divergence measure," *IEEE Transactions on Signal Processing*, vol. 64, no. 3, pp. 580–591, 2016.

[2] Herman Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *The Annals of Mathematical Statistics*, pp. 493–507, 1952.

[3] Guorong Xuan, Xiuming Zhu, Peiqi Chai, Zhenping Zhang, Yun Q Shi, and Dongdong Fu, "Feature selection based on the bhattacharyya distance," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*. IEEE, 2006, vol. 4, pp. 957–957.

[4] Ji-Gang Zhang and Hong-Wen Deng, "Gene selection for classification of microarray data based on the bayes error," *BMC bioinformatics*, vol. 8, no. 1, pp. 370, 2007.

[5] Constantino Carlos Reyes-Aldasoro and Abhir Bhalerao, "The bhattacharyya space for feature selection and its application to texture segmentation," *Pattern Recognition*, vol. 39, no. 5, pp. 812–826, 2006.

[6] Richard S Sutton and Andrew G Barto, *Reinforcement learning: An introduction*, vol. 1, MIT press Cambridge, 1998.

[7] Kevin Moon and Alfred Hero, "Multivariate f-divergence estimation with confidence," in *Advances in Neural Information Processing Systems*, 2014, pp. 2420–2428.

[8] Akshay Krishnamurthy, Kirthevasan Kandasamy, Barnabas Poczos, and Larry Wasserman, "Nonparametric estimation of renyi divergence and friends," *arXiv preprint arXiv:1402.2966*, 2014.

[9] Jillian Beardwood, John H Halton, and John Michael Hammersley, "The shortest path through many points," in *Math Proc Cambridge*. Cambridge Univ Press, 1959, vol. 55, pp. 299–327.

[10] Alfred O Hero, J Costa, and Bing Ma, "Asymptotic relations between minimal graphs and alpha-entropy," *Comm. and Sig. Proc. Lab.(CSPL), Dept. EECS, University of Michigan, Ann Arbor, Tech. Rep*, vol. 334, 2003.

[11] 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.

[12] Morteza Noshad, Kevin R Moon, Salimeh Yasaei Sekeh, and Alfred O Hero III, "Direct estimation of information divergence using nearest neighbor ratios," *arXiv preprint arXiv:1702.05222*, 2017.

[13] Norbert Henze and Mathew D Penrose, "On the multivariate runs test," *Annals of statistics*, pp. 290–298, 1999.

[14] Norbert Henze, "A multivariate two-sample test based on the number of nearest neighbor type coincidences," *The Annals of Statistics*, pp. 772–783, 1988.

[15] Kevin R Moon, Alfred O Hero, and B Véronique Delouille, "Meta learning of bounds on the bayes classifier error," in *Signal Processing and Signal Processing Education Workshop (SP/SPE), 2015 IEEE*. IEEE, 2015, pp. 13–18.

[16] Morteza Noshad Iranzad and Alfred O Hero III, "Rate-optimal meta learning of classification error," *arXiv preprint arXiv:1710.11315*, 2017.

[17] 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.

[18] Ananda L Freire, Guilherme A Barreto, Marcus Veloso, and Antonio T Varela, "Short-term memory mechanisms in neural network learning of robot navigation tasks: A case study," in *Robotics Symposium (LARS), 2009 6th Latin American*. IEEE, 2009, pp. 1–6.

[19] Ananda Freire et al, "UCI machine learning repository," 2010.