Optimal "Anti-Bayesian" Parametric Pattern Classification for the Exponential Family Using Order Statistics Criteria

A. Thomas and B. John Oommen*

School of Computer Science, Carleton University, Ottawa, Canada : K1S 5B6

Abstract. This paper reports some pioneering results in which optimal parametric classification is achieved in a counter-intuitive manner, quite opposed to the Bayesian paradigm. The paper, which builds on the results of [1], demonstrates (with both theoretical and experimental results) how this can be done for some distributions within the exponential family. To be more specific, within a Bayesian paradigm, if we are allowed to compare the testing sample with only *a single* point in the feature space from each class, the *optimal* Bayesian strategy would be to achieve this based on the (Mahalanobis) distance from the corresponding means, which in one sense, is the most *central* point in the respective distribution. In this paper, we shall show that we can obtain optimal results by operating in a diametrically opposite way, i.e., a so-called "anti-Bayesian" manner. Indeed, we shall show that by working with a *very few* (sometimes as small as two) points *distant* from the mean, one can obtain remarkable classification accuracies. These points, in turn, are determined by the *Order Statistics* of the distributions, and the accuracy of our method, referred to as Classification by Moments of Order Statistics (CMOS), attains the optimal Bayes' bound! In this paper, we shall show the claim for two uni-dimensional members of the exponential family. The theoretical results, which have been verified by rigorous experimental testing, also present a theoretical foundation for the families of Border Identification (BI) reported algorithms.

Keywords: Classification using Order Statistics (OS), Moments of OS.

1 Introduction

It is well known that when the expressions for the Bayesian classification (that involve maximizing the *a posteriori* probability) are simplified, this often reduces to testing the sample point using the corresponding distances/norms to the *means* or the "central points" of the distributions. Such a classification attains the Bayesian optimal lower bound.

* *Chancellor's Professor* ; *Fellow: IEEE* and *Fellow: IAPR*. This author is also an *Adjunct Professor* with the University of Agder in Grimstad, Norway. The work of this author was partially supported by NSERC, the Natural Sciences and Engineering Research Council of Canada.

A. Campilho and M. Kamel (Eds.): ICIAR 2012, Part I, LNCS 7324, pp. 11–18, 2012.

In this paper, in which we build on the results of [1], we shall demonstrate that we can obtain optimal results by operating in a diametrically opposite way, i.e., a so-called "anti-Bayesian" manner. Indeed, we shall show the completely counter-intuitive result that by working with a *few* points *distant* from the mean, one can obtain remarkable classification accuracies. The number of points referred to can be as small as *two* in the uni-dimensional case. Further, if these points are determined by the *Order Statistics* of the distributions, the accuracy attains the optimal Bayes' bound! Thus, put in a nut-shell, we introduce here the theory of optimal pattern classification using Order Statistics of the features rather than the distributions of the features themselves. Our novel methodology, is referred to as Classification by Moments of Order Statistics (CMOS), and this paper proves these results for two distributions within the exponential family.

The paper also formulates the theoretical rationale for the recently-developed families of Border Identification (BI) and some Prototype Reduction Schemes (PRS) algorithms [2,3]. In both these cases, instead of considering all the training patterns for the classification, a subset of the whole set is selected based on certain criteria. The learning (or training) is then performed on this reduced training set, which is also called the "Reference" set. This Reference set not only contains the patterns which are closer to the true discriminant's boundary, but also the patterns from the other regions of the space that can adequately represent the entire training set. However, in the interest of brevity, the details of BI and PRS algorithms are omitted here. They can be found in [4] and [5], where the parallels of these and our present results are explained in detail.

Contributions of this Paper: The novel contributions of this paper are:

- We propose an "anti-Bayesian" paradigm for the classification of patterns within the parametric mode of computation, where the distance computations are not with regard to the "mean" but with regard to some samples "distant" from the mean. These points, which are sometimes as few as *two*, are the moments of OS of the distributions;
- We provide a theoretical framework for adequately responding to the question of why the border points are more informative for classification;
- To justify these claims, we submit a formal analysis and the results of various experiments which have been performed for two distributions within the exponential family, and the results are clearly conclusive.

Our results for classification using the OS are both pioneering and novel.

2 Relevant Background Areas Regarding Order Statistics

Let $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$ be a univariate random sample of size n that follows a continuous distribution function Φ, where the probability density function (pdf) is $\varphi(\cdot)$. Let $\mathbf{x}_{1,n}$, $\mathbf{x}_{2,n}$, ..., $\mathbf{x}_{n,n}$ be the corresponding Order Statistics (OS). The r^{th} OS, $\mathbf{x}_{r,n}$, of the set is the r^{th} smallest value among the given random variables. The pdf of $\mathbf{y} = \mathbf{x}_{r,n}$ is given by:

$$f_{\mathbf{y}}(y) = \frac{n!}{(r-1)!(n-r)!} \{\Phi(y)\}^{r-1} \{1-\Phi(y)\}^{n-r} \varphi(y),$$

where $r = 1, 2, ..., n$. The reasoning for the above expression is straightforward. If the r^{th} OS appears at a location given by $\mathbf{y} = \mathbf{x}_{r,n}$, it implies that the $r-1$ smaller elements of the set are drawn independently from a Binomial distribution with a probability $\Phi(y)$, and the other $n-r$ samples are drawn using the probability $1-\Phi(y)$. The factorial terms result from the fact that the $(r-1)$ elements can be independently chosen from the set of n elements.

Using the distribution $f_{\mathbf{y}}(y)$, the k^{th} moment of $\mathbf{x}_{r,n}$ can be formulated as:

$$E[\mathbf{x}_{r,n}^k] = \frac{n!}{(r-1)!(n-r)!} \int_{-\infty}^{+\infty} y^k \Phi(y)^{k-1}(x)(1-\Phi(y))^{n-r}\varphi(y)dy,$$

provided that both sides of the equality exist [6,7].

The fundamental theorem concerning the OS that we invoke is found in many papers [7,8,9]. The theorem can be summarized as follows.

Let $n \geq r \geq k+1 \geq 2$ be integers. Then, since Φ is a nondecreasing and right-continuous function from $\mathbb{R} \rightarrow \mathbb{R}$, $\Phi(\mathbf{x}_{r,n})$ is uniform in [0,1]. If we now take the k^{th} moment of $\Phi(\mathbf{x}_{r,n})$, it has the form [8]:

$$E[\Phi^k(\mathbf{x}_{r,n})] = \frac{B(r+k, n-r+1)}{B(r, n-r+1)} = \frac{n!\ (r+k-1)!}{(n+k)!\ (r-1)!}, \tag{1}$$

where $B(a,b)$ denotes the *Beta* function, and $B(a,b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}$ since its parameters are integers.

The above fundamental result can also be used for characterization purposes as explained in [5,8]. The implications of the above are the following:

1. If $n = 1$, implying that only a *single* sample is drawn from $\mathbf{x}$, from Eq. (1),

$$E[\Phi^1(\mathbf{x}_{1,1})] = \frac{1}{2}, \implies E[\mathbf{x}_{1,1}] = \Phi^{-1}\left(\frac{1}{2}\right), \tag{2}$$

   which is the median of the distribution.
2. If $n = 2$, implying that only *two* samples are drawn from $\mathbf{x}$, we see that:

$$E[\Phi^1(\mathbf{x}_{1,2})] = \frac{1}{3}, \implies E[\mathbf{x}_{1,2}] = \Phi^{-1}\left(\frac{1}{3}\right), \text{and} \tag{3}$$

$$E[\Phi^1(\mathbf{x}_{2,2})] = \frac{2}{3}, \implies E[\mathbf{x}_{2,2}] = \Phi^{-1}\left(\frac{2}{3}\right). \tag{4}$$

   Thus, the first moment of the first and second 2-order OS would be the values where Φ equal $\frac{1}{3}$ and $\frac{2}{3}$ respectively.

3 Optimal Bayesian Classification Using 2-OS

3.1 The Generic Classifier

Having characterized the moments of the OS of arbitrary distributions, we shall now consider how they can be used to design a classifier.

Let us assume that we are dealing with the 2-class problem with classes ω_1 and ω_2, where their class-conditional densities are $f_1(x)$ and $f_2(x)$ respectively (i.e, their corresponding distributions are $F_1(x)$ and $F_2(x)$ respectively)[1]. Let ν_1 and ν_2 be the corresponding *medians* of the distributions. Then, classification based on ν_1 and ν_2 would be the strategy that classifies samples based on a *single* OS. We shall show the fairly straightforward result that for all symmetric distributions, this classification accuracy attains the Bayes' accuracy.

This result is not too astonishing because the median is centrally located close to (if not exactly) on the mean. The result for higher order OS is actually far more intriguing because the higher order OS are not located centrally (close to the means), but rather distant from the means. Consequently, we shall show that for a large number of distributions, mostly from the exponential family, the classification based on *these* OS again attains the Bayes' bound.

In [1], we initiated this discussion by examining the Uniform distribution. The reason for this was that even though the distribution itself is rather trivial, the analysis provided us with an insight into the mechanism by which the problem can be tackled, which can then be extended for other distributions. Here, we proceed to consider the CMOS for other distributions in the exponential family.

3.2 The Laplace (or Doubly-Exponential) Distribution

The *Laplace distribution* is a continuous uni-dimensional pdf named after Pierre-Simon Laplace. It is sometimes called the *doubly exponential distribution*, because it can be perceived as being a combination of two exponential distributions, with an additional location parameter, spliced together back-to-back.

If the densities of ω_1 and ω_2 are doubly exponentially distributed,

$$f_1(x) = \frac{\lambda_1}{2} e^{-\lambda_1 |x - c_1|}, \quad -\infty < x < \infty, \text{and}$$

$$f_2(x) = \frac{\lambda_2}{2} e^{-\lambda_2 |x - c_2|}, \quad -\infty < x < \infty,$$

where c_1 and c_2 are the respective means of the distributions. By elementary integration and straightforward algebraic simplifications, the variances of the distributions can be seen to be $\frac{2}{\lambda_1^2}$ and $\frac{2}{\lambda_2^2}$ respectively.

If $\lambda_1 \neq \lambda_2$, the samples can be classified based on the heights of the distributions and their point of intersection. The formal results for the general case are a little more complex. However, to initiate discussions, we shall first consider the case when $\lambda_1 = \lambda_2$. In this scenario, the reader should observe the following:

- Because the distributions have the equal height, i.e. $\lambda_1 = \lambda_2$, the testing sample $\mathbf{x}$ will obviously be assigned to ω_1 if it is less than c_1 and be assigned to ω_2 if it is greater than c_2.
- Further, the crucial case is when $c_1 < x < c_2$. In this regard, we shall analyze the CMOS classifier and prove that it attains the Bayes' bound even when one uses as few as *only* 2 OSs.

[1] Throughout this section, we will assume that the *a priori* probabilities are equal.

Theoretical Analysis: Doubly-Exponential Distribution - 2-OS. We shall first derive the moments of the 2-OS for the doubly exponential distribution. By virtue of Eq. (3) and (4), the expected values of the first moments of the two OS can be obtained by determining the points where the cumulative distribution function attains the values $\frac{1}{3}$ and $\frac{2}{3}$. Let u_1 be the point for the percentile $\frac{2}{3}$ of the first distribution, and u_2 be the point for the percentile $\frac{1}{3}$ of the second distribution. Then:

$$\int_{c_1}^{u_1} \frac{\lambda_1}{2}\, e^{-\lambda_1|x-c_1|}dx \;=\; \frac{2}{3} - \frac{1}{2} = \frac{1}{6}, \text{ and} \tag{5}$$

$$\int_{-\infty}^{u_2} \frac{\lambda_2}{2}\, e^{\lambda_2|x-c_2|}dx \;=\; \frac{1}{3}. \tag{6}$$

The points of interest, i.e., u_1 and u_2, can be obtained by straightforward integrations and simplifications as follows:

$$\int_{c_1}^{u_1} \frac{\lambda_1}{2}\, e^{-\lambda_1|x-c_1|}dx \;= \frac{1}{6} \implies u_1 = c_1 - \frac{1}{\lambda_1}\, log\left(\frac{2}{3}\right). \tag{7}$$

$$u_2 = c_2 + \frac{1}{\lambda_2}\, log\left(\frac{2}{3}\right). \tag{8}$$

With these points at hand, we shall now demonstrate that, for doubly exponential distributions, the classification based on the expected values of the moments of the 2-OS, CMOS, attains the Bayesian bound.

Theorem 1. *For the 2-class problem in which the two class conditional distributions are Doubly Exponential and identical, CMOS, the classification using two OS, attains the optimal Bayes' bound.*

Proof. This proof is omitted here in the interest of space. It is found in [4]. □

Experimental Results: Doubly-Exponential Distribution - 2OS. The CMOS classifier was rigorously tested for a number of experiments with various Doubly Exponential distributions having means c_1 and c_2. In every case, the 2-OS CMOS gave exactly the same classification as that of the Bayesian classifier. The method was executed 50 times with the 10-fold cross validation scheme. The test results are depicted in Table 1.

From the experimental results and the theoretical analysis, we conclude that the expected values of the first moment of the 2-OS of the Doubly Exponential distribution can always be utilized to yield the exact accuracy as that of the Bayes' bound, even though this is a drastically anti-Bayesian operation.

We now proceed to consider the analogous result for the k-OS.

Theoretical Analysis: Doubly-Exponential Distribution - k-OS. We now extend the results of Theorem 1 for the case when we utilize other k-OS for the CMOS. The formal result pertaining to this is given in Theorem 2.

Table 1. Classification for the Doubly Exponential Distribution by the CMOS

$\mathbf{c_1}$	0	0	0	0	0	0	0	0	0
$\mathbf{c_2}$	10	9	8	7	6	5	4	3	2
Bayesian	99.75	99.65	99.25	99.05	98.9	97.85	96.8	94.05	89.9
CMOS	99.75	99.65	99.25	99.05	98.9	97.85	96.8	94.05	89.9

Theorem 2. *For the* 2*-class problem in which the two class conditional distributions are Doubly Exponential and identical, the optimal Bayesian classification can be achieved by using symmetric pairs of the n-OS, i.e., the* $n-k$ *OS for* ω_1 *and the* k *OS for* ω_2 *if and only if* $log\left(\frac{2k}{n+1}\right) > \frac{c_1-c_2}{2}$.

Proof. The proof of this result is included in [4]. □

Experimental Results: Doubly-Exponential Distribution - k-OS
The CMOS method has been rigorously tested with different possibilities of k-OS and for various values of n, and the test results are given in Table 2.

Table 2. Results of the classification obtained by using the symmetric pairs of the OS for different values of n. The value of c_1 and c_2 were set to be 0 and 3.

No.	**Order**(n)	**Moments**	OS_1	OS_2	**CMOS**	**Pass/Fail**
1	Two	$\left(\frac{2}{3},\frac{1}{3}\right)$	$c_1 - \frac{1}{\lambda_1}log\left(\frac{2}{3}\right)$	$c_2 + \frac{1}{\lambda_2}log\left(\frac{2}{3}\right)$	95.2	Passed
2	Three	$\left(\frac{3}{4},\frac{1}{4}\right)$	$c_1 - \frac{1}{\lambda_1}log\left(\frac{1}{2}\right)$	$c_2 + \frac{1}{\lambda_2}log\left(\frac{1}{2}\right)$	95.2	Passed
3	Four	$\left(\frac{5-i}{5},\frac{i}{5}\right), 1 \le i \le \frac{n}{2}$	$c_1 - \frac{1}{\lambda_1}log\left(\frac{4}{5}\right)$	$c_2 + \frac{1}{\lambda_2}log\left(\frac{4}{5}\right)$	95.2	Passed
4	Five	$\left(\frac{6-i}{6},\frac{i}{6}\right), 1 \le i \le \frac{n}{2}$	$c_1 - \frac{1}{\lambda_1}log\left(\frac{1}{3}\right)$	$c_2 + \frac{1}{\lambda_2}log\left(\frac{1}{3}\right)$	95.2	Passed
5	Six	$\left(\frac{7-i}{7},\frac{i}{7}\right) 1 \le i \le \frac{n}{2}$	$c_1 - \frac{1}{\lambda_1}log\left(\frac{4}{7}\right)$	$c_2 + \frac{1}{\lambda_2}log\left(\frac{4}{7}\right)$	95.2	Passed
6	Seven	$\left(\frac{8-i}{8},\frac{i}{8}\right) 1 \le i \le \frac{n}{2}$	$c_1 - \frac{1}{\lambda_1}log\left(\frac{1}{4}\right)$	$c_2 + \frac{1}{\lambda_2}log\left(\frac{1}{4}\right)$	95.2	Passed
7	Eight	$\left(\frac{9-i}{9},\frac{i}{9}\right), 1 \le i \le \frac{n}{2}$	$c_1 - \frac{1}{\lambda_1}log\left(\frac{2}{9}\right)$	$c_2 + \frac{1}{\lambda_2}log\left(\frac{2}{9}\right)$	4.8	Failed
8	Eight	$\left(\frac{9-i}{9},\frac{i}{9}\right), 1 \le i \le \frac{n}{2}$	$c_1 - \frac{1}{\lambda_1}log\left(\frac{4}{9}\right)$	$c_2 + \frac{1}{\lambda_2}log\left(\frac{4}{9}\right)$	95.2	Passed
9	Nine	$\left(\frac{10-i}{10},\frac{i}{10}\right), 1 \le i \le \frac{n}{2}$	$c_1 - \frac{1}{\lambda_1}log\left(\frac{3}{5}\right)$	$c_2 + \frac{1}{\lambda_2}log\left(\frac{3}{5}\right)$	95.2	Passed

To clarify the table, consider the row given by Trial No. 5 in which the 6-OS were invoked for the classification. In this case, the possible symmetric OS pairs could be $\langle 1,6\rangle$, $\langle 2,5\rangle$, and $\langle 3,4\rangle$ respectively. Observe that the expected values for the first moment of the k-OS has the form $E[\mathbf{x}_{k,n}] = log\left(\frac{2k}{n+1}\right)$. In every single case, the accuracy attained the Bayes' bound, as seen in the table.

Now, consider the results presented in the row denoted by Trial No. 7. In this case, the testing attained the Bayes accuracy for the symmetric OS pairs $\langle 2,7\rangle$, $\langle 3,6\rangle$ and $\langle 4,5\rangle$ respectively. However, the classifier "failed" for the specific 8-OS, when the OS used were $c_1 - \frac{1}{\lambda_1}log\left(\frac{2}{9}\right)$ and $c_2 + \frac{1}{\lambda_2}log\left(\frac{2}{9}\right)$, as these values violate the condition $log\left(\frac{2k}{n+1}\right) > \frac{c_1-c_2}{2}$, imposed by Theorem 2. Observe that if $log\left(\frac{2k}{n+1}\right) < \frac{c_1-c_2}{2}$, the symmetric pairs should be reversed to obtain optimality.

The multi-dimensional case is currently being investigated and will be published in a forthcoming paper.

3.3 The Gaussian Distribution

The Normal (or Gaussian) distribution is a continuous probability distribution that is often used as a first approximation to describe real-valued random variables that tend to cluster around a single mean value. It is particularly pertinent due to the so-called Central Limit Theorem. The distribution's pdf is:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}.$$

Theoretical Analysis: Gaussian Distribution
Working with the OS of Normal distributions is extremely cumbersome because its density function is not integrable in a closed form. One has to resort to tabulated cumulative error functions or to numerical methods to obtain precise percentile values. However, a lot of work has been done in this area for *certain* OS, and can be found in [6,8,10,11], from which we can make some interesting conclusions.

The moments of the OS for the Normal distribution can be determined from the generalized expression:

$$E[\mathbf{x}_{k,n}^r] = \frac{n!}{(k-1)!(n-k)!} \int_{-\infty}^{+\infty} x^r \Phi^{k-1}(x)(1-\Phi(x))^{n-k}\varphi(x)dx,$$

where $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$ and $\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt$. From this expression, the expected values of the first moment of the 2-OS can be determined as $E[\mathbf{x}_{1,2}] = \mu - \frac{\sigma}{\sqrt{2\pi}}$ and $E[\mathbf{x}_{2,2}] = \mu + \frac{\sigma}{\sqrt{2\pi}}$ as shown in [6]. Using this, we now show that the CMOS with 2-OS yields the same Bayesian accuracy.

Theorem 3. *For the 2-class problem in which the two class conditional distributions are Gaussian and identical, CMOS, the classification using 2-OS, attains the optimal Bayes' bound.*

Proof. The proof of this theorem is omitted here but found in [4]. □

Experimental Results: Gaussian Distribution
After the data points were generated, the CMOS classifier was rigorously tested for a number of experiments with various Gaussian distributions having means μ_1 and μ_2. In every case, the 2-OS CMOS gave *exactly* the same accuracy as that of the Bayesian classifier. The method was executed 50 times with the 10-fold cross validation scheme. The test results are displayed in Table 3, whence the power of the scheme is clear!

We believe that the optimal Bayes' bound can also be attained by performing the classification with respect to the k-OS. However, as the density function is not integrable, the expected values of the moments of the k-OS should rather be obtained by numerical integration, and is currently being done. The classification for the multi-dimensional classes is currently being investigated.

Table 3. Classification of Normally distributed classes by the CMOS 2-OS method for different means

μ_1	0	0	0	0	0	0
μ_2	14	12	10	8	6	4
Bayesian	99.2	96.5	95.1	95	90	85
CMOS	99.2	96.5	95.1	95	90	85

4 Conclusions

In this paper, we have shown that optimal classification can be attained by an "anti-Bayesian" approach, i.e., by working with a *very few* (sometimes as small as two) points *distant* from the mean. This scheme, referred to as CMOS, Classification by Moments of Order Statistics, operates by using these points determined by the *Order Statistics* of the distributions. In this paper, which has built on the results of [1], we have proven the claim for two uni-dimensional distributions within the exponential family, and the theoretical results have been verified by rigorous experimental testing. Our results for classification using the OS are both pioneering and novel.

References

1. Thomas, A., Oommen, B.J.: Optimal "Anti-Bayesian" Parametric Pattern Classification Using Order Statistics Criteria (accepted for Publication, 2012)
2. Garcia, S., Derrac, J., Cano, J.R., Herrera, F.: Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study. IEEE Transactions on Pattern Analysis and Machine Intelligence
3. Triguero, I., Derrac, J., Garcia, S., Herrera, F.: A Taxonomy and Experimental Study on Prototype Generation for Nearest Neighbor Classification. IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews
4. Thomas, A., Oommen, B.J.: The Foundational Theory of Optimal "Anti-Bayesian" Parametric Pattern Classification Using Order Statistics Criteria (to be submitted, 2012)
5. Thomas, A.: Pattern Classification using Novel Order Statistics and Border Identification Methods. PhD thesis, School of Computer Science, Carleton University (to be submitted, 2013)
6. Ahsanullah, M., Nevzorov, V.B.: Order Statistics: Examples and Exercises. Nova Science Publishers, Inc. (2005)
7. Morris, K.W., Szynal, D.: A goodness-of-fit for the Uniform Distribution based on a Characterization. Journal of Mathematical Science 106, 2719–2724 (2001)
8. Lin, G.D.: Characterizations of Continuous Distributions via Expected values of two functions of Order Statistics. Sankhya: The Indian Journal of Statistics 52, 84–90 (1990)
9. Too, Y., Lin, G.D.: Characterizations of Uniform and Exponential Distributions. Academia Sinica 7(5), 357–359 (1989)
10. Grudzien, Z., Szynal, D.: Characterizations of Distributions by Moments of Order Statistics when the Sample Size is Random. Applications Mathematicae 23, 305–318 (1995)
11. Nadarajah, S.: Explicit Expressions for Moments of Order Statistics. Statistics and Probability Letters 78, 196–205 (2008)