A New Paradigm for Pattern Classification: Nearest *Border* Techniques

Yifeng Li[1], B. John Oommen[2], Alioune Ngom[1], and Luis Rueda[1]

[1] School of Computer Science, University of Windsor, Canada
{li11112c,angom,lrueda}@uwindsor.ca
[2] School of Computer Science, Carleton University, Canada
Also *Adjunct Professor*, University of Agder, Grimstad, Norway
oommen@scs.carleton.ca

Abstract. There are many paradigms for pattern classification. As opposed to these, this paper introduces a paradigm that has not been reported in the literature earlier, which we shall refer to as the Nearest *Border* (NB) paradigm. The philosophy for developing such a NB strategy is as follows: Given the training data set for each class, we shall first attempt to create borders for each individual class. After that, we advocate that testing is accomplished by assigning the test sample to the class *whose border it lies closest to*. This claim is actually counter-intuitive, because unlike the centroid or the median, these border samples are often "outliers" and are, really, the points that represent the class the least. However, we have formally proven this claim, and the theoretical results have been verified by rigorous experimental testing.

Keywords: Pattern Classification, Border Identification, SVM.

1 Introduction

The problem of classification in machine learning can be quite simply described as follows: If we are given a limited number of training samples, and if the class-conditional distributions are unknown, the task at hand is to predict the class label of a new sample with minimum risk. Within the generative model, one resorts to modeling the class-conditional distributions $p(\boldsymbol{x}|w_i)$ and priors $p(w_i)$ and $p(\boldsymbol{x})$, and then computing the *a posteriori* distribution $p(w_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|w_i)p(w_i)}{p(\boldsymbol{x})}$ after the testing sample arrives. The strength of this strategy is that one obtains an optimal performance if the assumed distributions are the same as the actual one. The limitation, of course, is that it is often difficult, if not impossible, to compute. The alternative is to directly approximate the posterior distribution itself. This paper advocates such a philosophy.

The goal of this paper is to present a new paradigm in pattern recognition, which we shall refer to as the Nearest *Border* (NB) paradigm. This archetype possesses similarities to many of the well-established methodologies in pattern recognition, and can also be seen to include many of *their* salient facets/traits.

S. Cranefield and A. Nayak (Eds.): AI 2013, LNAI 8272, pp. 441–446, 2013.

There are four family algorithms that are most closely related to our NB paradigm. They include i) Prototype Reduction (PR) schemes [6], ii) Border Identification (BI) algorithms [6], iii) "Anti-Bayesian" Order-Statistics (OS) based algorithms [6], and iv) Support Vector Machines (SVMs) [8].

The novel contributions of this paper are the following:

1. We propose a new pattern recognition paradigm, the Nearest *Border* paradigm, in which we create borders for each individual class, and where testing is accomplished by assigning the test sample to the class whose border it lies closest to.
2. Our paradigm falls within the family of PR schemes, because it yields a reference set which is a small subset of original training patterns. The testing is achieved by *only* utilizing the latter.
3. Our paradigm falls within the family of BI methods.
4. The Nearest *Border* paradigm is essentially "anti-Bayesian" in its salient characteristics. This is because the testing is not done based on central concepts such as the centroid or the median, but by comparisons using these border samples, which are often "outliers" and which, in one sense, represent the class the least.
5. The Nearest *Border* paradigm is closely related to the family of SVMs, because the computations and optimization used are similar to those involved in deriving SVMs.

2 Method

2.1 The Theory of the Nearest Border (NB) Paradigm

We assume that we are dealing with a classification problem involving g classes: $\{\omega_1, \cdots, \omega_g\}$. For any specific class ω_i, we define a region $\mathcal{R}_i$ that is described by the function $f_i(\boldsymbol{x}) = 0$ (which we shall refer to as its "border"), where $\mathcal{R}_i = \{\boldsymbol{x} | f_i(\boldsymbol{x}) > 0\}$. We describe $\mathcal{R}_i$ in this manner so that it is able to capture the main mass of the probability distribution $p_i(\boldsymbol{x}) = p(\boldsymbol{x}|\omega_i)$. All points that lie outside of $\mathcal{R}_i$, are said to fall in its "outer" region, $\bar{\mathcal{R}}_i$, where $\bar{\mathcal{R}}_i = \{\boldsymbol{x} | f_i(\boldsymbol{x}) < 0\}$. These points are treated as outliers as far as class ω_i is concerned. The function $f_i(\boldsymbol{x})$ is crucial to our technique because it explicitly defines the region $\mathcal{R}_i$. Formally, the function $f_i(\boldsymbol{x})$ must be defined in such a way that:

1. $f_i(\boldsymbol{x})$ is the *signed distance* from the point $\boldsymbol{x}$ to the border such that $f_i(\boldsymbol{x}) > 0$ if $\boldsymbol{x} \in \mathcal{R}_i$, and $f_i(\boldsymbol{x}) < 0$ if $\boldsymbol{x} \in \bar{\mathcal{R}}_i$;
2. If $f_i(\boldsymbol{x}_1) > f_i(\boldsymbol{x}_2)$, then $p_i(\boldsymbol{x}_1) > p_i(\boldsymbol{x}_2)$;
3. If $f_i(\boldsymbol{x}) > f_j(\boldsymbol{x})$, then $p(w_i|\boldsymbol{x}) > p(w_j|\boldsymbol{x})$.

In order to predict the class label of a new sample $\boldsymbol{x}$, we calculate its signed distance from each class, and thereafter assign it to the class with the minimum distance. In other words, we invoke the softmax rule: $j = \arg\max_{i=1}^{g} f_i(\boldsymbol{x})$.

The main challenge that we face in formulating, designing and implementing such a NB theory lies in the complexity of conveniently and accurately procuring

such borders. The reader will easily see that this is equivalent to the problem of identifying functions $\{f_i(\boldsymbol{x})\}$ that satisfy the above constraints. Although a host of methods to do this are possible, in this paper, we propose one that identifies the boundaries using the one-class SVM.

2.2 NB Classifiers: The Implementations of the NB Paradigm

The basic Nearest Centroid (NC) approach only uses the means of the class-conditional distribution, and this is the reason why it is not effective for the scenario when the variances of the various classes are very different. The NC scheme can be extended to allow different class variance by using, for example, Gaussian Mixture Model. The difficulty of extending any linear model, e.g. SVM, from its two-class formulation to its corresponding multi-class formulation, lies in the fact that a hyperplane always partitions the feature space into two "open" subspaces, implying that this can lead to ambiguous regions that may be generated by some extensions of the two-class regions for the multi-class case. The most popular schemes to resolve this are the one-against-all (using a softmax function) and one-against-one solutions.

As an one-class model, the work based on Tax and Duin's Support Vector Domain Description (SVDD or one-class SVM) [5] aims to find a closed hypersphere in the feature space that captures the main part of the distribution. By examining the corresponding SVM, we see that the hypersphere obtained by the SVDD is the estimate of the features' *Highest Density Region* (HDR). In particular, for the univariate distribution, the estimation of the *Highest Density Interval* (HDI) involves searching for the threshold p^* that satisfies: $\int_{x:p(x|D)>p^*} p(x|D)dx = 1-\alpha$. The $(1-\alpha)\%$ HDI is defined as $C_\alpha(p^*) = \{x : p(x|D) \geq p^*\}$. If we now define the *Central Interval* (CI) by the interval:

$$C_\alpha(l,u) = \{x \in (l,u) | P(l \leq x \leq u|D) = 1-\alpha, P(x \leq l) = \frac{\alpha}{2}, P(x \geq u) = \frac{\alpha}{2}\},$$

one will see that, for symmetric unimodal univariate distribution, HDI coincides with the CI. However, for nonsymmetric univariate distributions, the HDI is smaller than the CI. For known distributions, the CI can be estimated by the corresponding quantile. However, for unknown distributions, the CI can be estimated by a Monte Carlo approximation (or by the histogram, or the *Order Statistics*). However, in the context of this paper, we remark that by virtue of Vapnik's principle, it is not necessary to estimate the density by invoking a non-parametric method. For multivariate distributions, we can estimate the $(1-\alpha)\%$ HDR $C_\alpha(f)$ by using the equation:

$$\min_f \int_{f(x)\geq 0} 1dx, \text{ s.t. } \int_{x:f(x)\geq 0} p(x|D)dx = 1-\alpha. \tag{1}$$

We shall refer to this optimal contour $f^*(x) = 0$ as the $(1-\alpha)$-border/contour.

Our idea for classification is the following: We can learn a hypersphere using SVDD for each class in the feature space in order to describe the border of this

class. We then calculate the distance from a unknown sample to the border of each class and assign it to the class with the minimum distance. The training phase of our approach is to learn the hypersphere $f_i(\boldsymbol{x}) = 0$ for each class. The prediction phase then involves assigning the unknown sample $\boldsymbol{x}$ using the rule: $j = \arg\max_{i=1}^{g} f_i(\boldsymbol{x})$. In particular, we note that:

1. $f_i(\boldsymbol{x}) \in \mathbb{R}$ is the signed distance of $\boldsymbol{x}$ from the corresponding boundary;
2. For points inside the i^{th} hypersphere, $f_i(\boldsymbol{x}) > 0$;
3. For points outside the hypersphere, $f_i(\boldsymbol{x}) < 0$. Further, the larger $f_i(\boldsymbol{x})$ is, the closer it is to class ω_i, and the higher the value of $p(w_i|\boldsymbol{x})$ is. From the parameters of $f_i(\boldsymbol{x})$, we can see that $f_i(\boldsymbol{x})$ considers both mean and variance of the distribution. It can be further enhanced by the *normalized distance* through the operation of dividing it by R_i (the radius of the hypersphere), that is $\frac{f_i(\boldsymbol{x})}{R_i}$.

We refer to this approach as the *Nearest Border approach based on Hyper-Sphere* (NB-HS). Hereafter, the hypersphere based NB using the un-normalized and normalized decision rules will be denoted by *ν-NB*, and *ν-NBN*, respectively, where ν is the upper bound of the fraction of outliers and the lower bound of the fraction of the support vectors in SVDD. As the number of training samples increases to infinity, these two bounds converge to ν. However, in practice, we usually have a very limited number of training samples. In order to obtain ν which corresponds to the α fraction of outliers, firstly, we need to let $\nu = \alpha$, and then reduce ν gradually until the α fraction of outliers are obtained. This variant of NB will be named the *α-NB* in the subsequent sections.

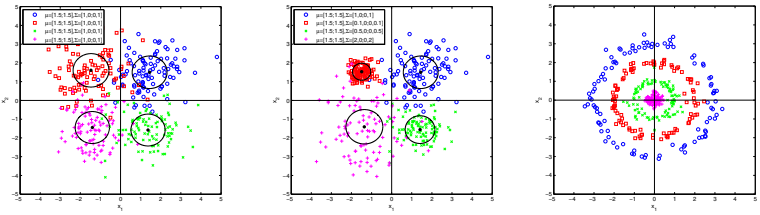
3 Experimental Results

The NB schemes were rigorously tested. Our computational experiments can be divided into two segments. First, we verify the capability of our method on three artificial data sets. Then, we statistically compared our approach with benchmark classifiers on 17 well-known real-life data sets.

Accuracy on Synthetic Data: We verified our methods on three synthetic data sets described as follows and shown in Fig. 1. Each data set has four classes and 100 two-dimensional points in each class. In the *SameVar* data, all classes have the same variance, while in *DiffVar*, the classes have different variances. *NonLinear* is a nonlinear data set.

For the artificial data, we compared our method with the Naive Bayes [2], 1-NN [2], NC [7], and SVM [4] classifiers. Linear kernel was used for the NBs, NC, and SVM on the first two data sets, and the *Radial Basis Function* (RBF) kernel was used on the last one. We ran a 3-fold cross-validation on each data set 20 times. The mean accuracies and standard deviations are shown in Fig. 2a.

On the *SameVar* data, first, we can see that there is no significant difference between the ν-NB and ν-NBN, and α-NB. All of them yielded an almost-equivalent accuracy as the Naive Bayes. Second, it can be seen from Fig. 1a that the NB was able to identify the centers of each class accurately. The borders



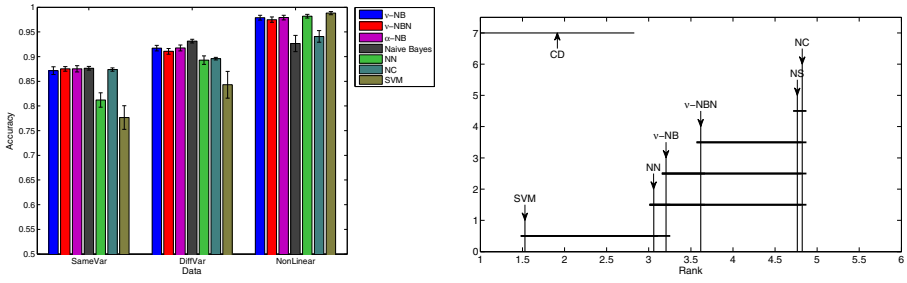
(a) SameVar **(b)** DiffVar **(c)** NonLinear

Fig. 1. Plots of the synthetic data sets

have the same volume, which demonstrates that the NB can identify the borders consistent with the variances. The NB approaches yielded an accuracy similar to the NC, which is reasonable because the identical variance of all classes is of no consequence to the NB. Third, the NN and SVM do not obtain comparable results, because the distance measure of the NN is affected by noise, and the SVM is not able to "disentangle" each class well using an one-versus-all scheme.

On the *DiffVar* data, first, we see that the results again confirm that the NB can identify the borders consistent with the variances (see Fig. 1b). The mean accuracies of all the NB approaches were very close to the Naive Bayes classifier. However, the NC yielded worse results than the NB. This is because the variance information helped the NB, while the NC scheme did not consider it.

Finally, for the *NonLinear* data, first, we affirm that all our NB methods and the SVM yielded comparably good results. Second, the Naive Bayes did not work well this time, because the data is not Gaussian. Further, the kernel NC was not competent either, because the data in the high-dimensional feature space have different variances for all the classes.



(a) Mean Performance On Synthetic Data **(b)** CD Diagram On Real-Life Data

Fig. 2. Performance on synthetic and real-life data

Statistical Test on Real-Life Data: In order to test the performance of our NB methods, we compared them with benchmark classifiers on 17 real-life data sets. The benchmark methods included were the 1-NN, NC, Nearest Subspace (NS) [3], and the SVM. We used the RBF kernel in our classifiers. We applied the Friedman test with Nemenyi test as a post-hoc test [1] on the accuracies of 3-fold

cross-validation. We set the significance level to $\alpha = 0.05$. Consequently the null hypothesis (all classifiers are equivalent) was rejected. The Crucial-Difference (CD) diagram of the Nemenyi test is illustrated in Fig. 2b.

First, as can be seen from the results, the difference between the ν-NB and the ν-NBN is negligible. However, ν-NB has a marginally higher rank than the ν-NBN. Second, the SVM obtained the highest rank. However, there is no significant difference among the SVM, the NN, and the ν-NB under the current significant level. This is quite a remarkable conclusion. Third, the performances of NC and NS are very close. Last, if we examine the accuracies of the classifiers, we can clearly identify two distinct groups: {SVM, NN, ν-NB, ν-NBN}, and {NC, NS}, demonstrating that our newly-introduced NB schemes are competitive to the best reported algorithms in the literature.

4 Conclusions and Future Work

In this paper, we introduced a new paradigm for classification which has not been reported in the literature. We refer to it as the Nearest *Border* paradigm. We emphasize that our methodology is actually counter-intuitive, because unlike the centroid or the median, these border samples are often "outliers" and are, indeed, the points that represent the class the least. The theoretical results have been verified by rigorous experimental testing. We preliminarily assume that the class-conditional distribution is unimodal and homoscedastic in feature space. We will focus on a method which is able to learn the border of complex distributions, for example using hyperellipse, local learning, or mixture models.

Acknowledgments. We acknowledge the valuable suggestions from the reviewers. This research is support by Canadian NSERC Grants #RGPIN228117-2011.

References

1. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30 (2006)
2. Mitchell, T.: Machine Learning. McGraw Hill, Ohio (1997)
3. Naseem, I., Togneri, R., Bennamoun, M.: Linear regression for face recognition. PAMI 32(11), 2106–2112 (2010)
4. Scholkopf, B., Smola, A., Williamson, B., Bartlett, P.: New support vector algorithm. Neural Computation 12, 1207–1245 (2000)
5. Tax, D., Duin, R.: Support vector domain description. Pattern Recognition Letters 20, 1191–1199 (1999)
6. Thomas, A., Oommen, B.J.: The fundamental theory of optimal "anti-Bayesian" parametric pattern classification using order statistics criteria. Pattern Recognition 46, 376–388 (2013)
7. Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G.: Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Statistical Science 18(1), 104–117 (2003)
8. Vapnik, V.: Statistical Learning Theory. Wiley-IEEE Press, New York (1998)