A Novel Border Identification Algorithm Based on an "Anti-Bayesian" Paradigm*

Anu Thomas and B. John Oommen**

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

Abstract. Border Identification (BI) algorithms, a subset of Prototype Reduction Schemes (PRS) aim to reduce the number of training vectors so that the reduced set (the border set) contains only those patterns which lie near the border of the classes, and have sufficient information to perform a meaningful classification. However, one can see that the true border patterns ("near" border) are not able to perform the task independently as they are not able to always distinguish the testing samples. Thus, researchers have worked on this issue so as to find a way to strengthen the "border" set. A recent development in this field tries to add more border patterns, i.e., the "far" borders, to the border set, and this process continues until it reaches a stage at which the classification accuracy no longer increases. In this case, the cardinality of the border set is relatively high. In this paper, we aim to design a novel BI algorithm based on a new definition for the term "border". We opt to select the patterns which lie at the border of the alternate class as the border patterns. Thus, those patterns which are neither on the true discriminant nor too close to the central position of the distributions, are added to the "border" set. The border patterns, which are *very small* in number (for example, five from both classes), selected in this manner, have the potential to perform a classification which is comparable to that obtained by well-known traditional classifiers like the SVM, and very close to the optimal Bayes' bound.

1 Introduction

The objective of a PRS is to reduce the cardinality of the training set to be as small as possible by selecting some training patterns based on various criteria, as long as the reduction does not significantly affect the performance. Thus, 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.

* The authors are grateful for the partial support provided by NSERC, the Natural Sciences and Engineering Research Council of Canada.

** *Chancellor's Professor* ; *Fellow: IEEE* and *Fellow: IAPR*. This author is also an *Adjunct Professor* with the University of Agder in Grimstad, Norway.

R. Wilson et al. (Eds.): CAIP 2013, Part I, LNCS 8047, pp. 196–203, 2013.
© Springer-Verlag Berlin Heidelberg 2013

Border Identification (BI) algorithms, which are a subset of PRSs, work with a Reference set which only contains "border" points. Specializing this criterion, the current-day BI algorithms, designed by Duch [1], Foody [2,3], and Li *et al.* [4], attempt to select a Reference set which contains border patterns derived, in turn, from the set of training patterns. Observe that, in effect, these algorithms also yield reduced training sets. Once the Reference set is obtained, all of these traditional methods perform the classification by invoking some sort of classifier like the SVM, neural networks etc. As opposed to the latter, we are interested in determining border patterns which, in some sense, are neither closer to the true optimal classifier nor to the means, and which can thus better classify the entire training set. Contrary to a Bayesian intuition, these border patterns have the ability to accurately classify the testing patterns, as we shall presently demonstrate. Our method is a combination of NN computations and (Mahalanobis) *multi*-dimensional[1] distance computations which yield the border points that are subsequently used for the purpose of classification. The characterizing component of our algorithm, referred to as ABBI, is that classification can be done by processing the obtained border points by themselves without invoking, for example, a subsequent SVM phase.

How then can one determine the border points themselves? This, indeed, depends on the model of computation - for example, whether we are working within the parametric or non-parametric model. The current paper deals with the former model, where the information about the classes is crystallized in the class-conditional *distributions* and their respective parameters, where the training samples are used to estimate the *parameters* of these models. In this paper, we have shown how the border points can be obtained by utilizing the information gleaned from the estimated distributions. Observe that with regard to classification and testing, all of these computations can be considered to be of a "pre-processing" nature, and so the final scheme would merely be of a Nearest Neighbor(NN)-sort. The details of how this is achieved is described in the paper.

2 A Novel Two-Class "Anti-Bayesian" BI Scheme

The Formal Algorithm. The problem of determining the border points for the parametric model of computation can be solved for fairly complex scenarios. When one examines the existing BI schemes, he observes that the information that has been utilized to procure the border patterns is primarily (and indeed, essentially) distance-based. In other words, the distances between the patterns are evaluated independently, and the border patterns are obtained based on such distances. The patterns obtained in this manner are considered as the new training set, which reduces these BI schemes to be special types of PRSs, but with the border patterns being the Reference set. However, as these border patterns

[1] We also have some initial results in which the distance and optimizations are done using lower-dimensional projections, the results of which are subsequently fused using an appropriate fusion technique.

are only the "Near" ones, they do not possess sufficient information to train an efficient classifier. We shall now rectify this.

We now mention a second major handicap associated with the traditional BI schemes. Once they have computed the border points associated with the specific classes, the traditional schemes operate by determining a "classifier" based on the new set. In other words, they have to determine a classifying *boundary* (linear, quadratic or SVM-based) to achieve this classification. As the reader will observe, in our work, we attempt to circumvent this entire phase. Indeed, in our proposed strategy, we merely achieve the classification using the final *small* subset of border points – which entails a significant reduction in computation.

The reader should also observe that this final decision would involve NN-like computations with a *few* points. The intriguing feature of these few points is that they lie close to the boundary and not to the mean, implying an "anti-Bayesian" philosophy [5,6,7].

In order to obtain the border patterns of the distributions ω_1 and ω_2 in an "anti-Bayesian" approach, we make use of the axiom that the patterns that have nearest neighbors from *other* classes *along* with the patterns of the same class fall into a common region - which is, by definition, the overlapped region.

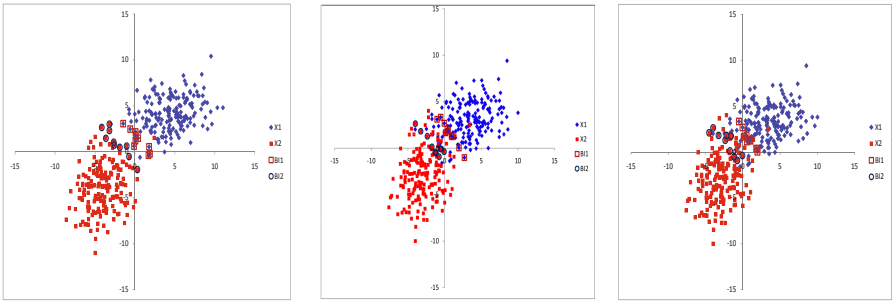
The proposed algorithm has 4 parameters, namely, J, J_1, J_2 and K. First of all, J denotes the number of border points that have to be selected from each class. We understand that in the process of selecting the border points, the training set must be "examined" so as to ignore the patterns which are not relevant for the classification. As this decision is taken based on the border points in and of themselves, we conclude that the patterns which are in the overlapping region are not able to provide an accurate decision, and so these points have to be ignored. Thus, for any X, those patterns with J_2 or more NNs out of the J_1 NNs, which are not from the same class as X, are ignored.

To be more specific, in order to eliminate the overlapping points, we first determine J_1-NNs of every pattern X. If J_2 or more of these NN patterns are from the same class, this pattern X is added to the new training set. Once this step is achieved, we are left with the training points which are not overlapping with any other classes. Thereafter, we evaluate the (Mahalanobis) distance[2](MD) of every pattern of the new training set with respect to the *mean* of both the classes. Both of these phases distinguish our particular strategy. The patterns which are almost equidistant from both the classes, and which are not determined to be overlapping with respect to the other classes, are added to the Border set.

The process of determining the (Mahalanobis) distances with respect to both the classes, is repeated for all the patterns of the new training set, and a decision is made for each pattern based on the difference between these distances.

The two-dimensional view of this philosophy is depicted in Figures 1a - 1c. The border patterns obtained by applying this method are also given in the figure, where the border patterns of class ω_1 are specified by rectangles, and those of class ω_2 are specified by circles. We now make the following observations:

[2] Any well-defined norm, appropriate for the data distribution, can be used to quantify this distance.



(a) Almost separable classes (b) Semi-overlapped classes (c) Overlapped classes

Fig. 1. Border patterns for separable and overlapped classes

1. If we examine Figure 1a, we can see that the border patterns that are specified by rectangles and circles are precisely those that lie at the true borders of the classes.
2. However, if the classes are semi-overlapped, then the "more interior" symmetric percentiles, such as the $\langle \frac{2}{3}, \frac{1}{3} \rangle$ can perform a near-optimal classification. This can be seen in Figure 1b. The patterns in this figure have more overlap (the $BD = 1.69$), and the border points chosen are the ones which lie just outside the overlapping region.
3. The same argument is valid for Figure 1c. In the OS-based classification, we have seen that if the classes have a large overlap as in Figure 1c (in this case, $MD = 0.78$), the border patterns again lie outside the overlapped region.

The algorithm for obtaining the border patterns, ABBI, is formally given in Algorithm 1.

Contrary to the traditional BI algorithms, ABBI requires only a *small* number of border patterns for the classification. For example, consider the Breast Cancer data set which contains 699 patterns. A traditional BI algorithms will obtain a border set of around 150 patterns for this data set. Furthermore, once these methods have obtained the border points, they will have to generate a classifier for the new reduced set to achieve the classification. As opposed to this, our method requires only 20 border patterns, and the classification is based on the *five* NN border patterns of the testing pattern.

3 Experimental Results

The proposed method ABBI has been tested on various data sets that include artificial and real-life data sets obtained from the UCI repository [8]. ABBI has also been compared with other well-known methods which include the NB, SVM, and the kNN. In order to obtain the results, ABBI algorithm was executed 50 times with the 10-fold cross validation scheme.

Algorithm 1. ABBI(ω_1, ω_2)

Input:

Data from two classes; ω_1, ω_2, whose means are $\mathbf{M}_1$ and $\mathbf{M}_2$ respectively.
Parameters: J_1, J, J_2, K: Small numbers

Assumption:

Dist computes the distance between two vectors.
DistDiff computes the difference in distances obtained with respect to μ_1 and μ_2

Notation:

NTR_1 and NTR_2 are the new training sets which do not contain points in the overlapped region.

Output:

The classification based only on the Border points

Method:

1: $NTR_1 \leftarrow \emptyset$
2: $NTR_2 \leftarrow \emptyset$
3: Divide points of ω_1 into training and testing sets, TRP_1 and T_1 respectively
4: Divide points of ω_2 into training and testing sets, TRP_2 and T_2 respectively
5: **for all** $X \in TRP_1$ **do**
6: Compute J_1 NNs of X
7: If J_2 or more NNs are from class ω_1, $NTR_1 \leftarrow NTR_1 \cup X$
8: **end for**
9: **for all** $X \in TRP_2$ **do**
10: Compute J_1 NNs of X
11: If J_2 or more NNs are from class ω_2, $NTR_2 \leftarrow NTR_2 \cup X$
12: **end for**
13: **for all** $X \in NTR_1$ **do**
14: Dist(X, M_1)
15: Dist(X, M_2)
16: **end for**
17: **for all** $X \in NTR_2$ **do**
18: Dist(X, M_1)
19: Dist(X, M_2)
20: **end for**
21: **for all** $X \in NTR_1$ **do**
22: DistDiff(X)
23: **end for**
24: **for all** $X \in NTR_2$ **do**
25: DistDiff(X)
26: **end for**
27: Add J points with minimum DistDiff from NTR_1 and NTR_2 to BI
28: Classify testing points using a K-NN based on the points in BI.

End Algorithm

Artificial Data Sets: For a *prima facie* testing of artificial data, we generated two classes that obeyed Gaussian distributions. To do this, we made use of a Uniform $[0, 1]$ random variable generator to generate data values that follow a Gaussian distribution. The expression $\mathbf{z} = \sqrt{-2ln(u_1)}\ cos(2\pi u_2)$ is known to yield data values that follow $N(0, 1)$ [9]. Thereafter, by using the technique described in [10], one can generate Gaussian random vectors which possess any arbitrary mean and covariance matrix. In our experiments, since this is just for a *prima facie* case, we opted to perform experiments for two-dimensional and three-dimensional data sets. The respective means of the classes were $[\mu_{11}, \mu_{12}]^T$ and $[\mu_{21}, \mu_{22}]^T$ for the two-dimensional data, and $[\mu_{11}, \mu_{12}, \mu_{13}]^T$ and $[\mu_{21}, \mu_{22}, \mu_{23}]^T$

for the three-dimensional data. Further, the corresponding covariance matrices of the two-dimensional classes had the forms:

$$\Sigma_1 = \begin{bmatrix} a^2 & \alpha ab \\ \alpha ab & b^2 \end{bmatrix}, \qquad \Sigma_2 = \begin{bmatrix} b^2 & \alpha ab \\ \alpha ab & a^2 \end{bmatrix}$$

The covariance matrices for the three-dimensional classes had the forms:

$$\Sigma_1 = \begin{bmatrix} a^2 & 0 & \alpha ab \\ 0 & 1 & 0 \\ \alpha ab & 0 & b^2 \end{bmatrix}, \qquad \Sigma_1 = \begin{bmatrix} b^2 & 0 & \alpha ab \\ 0 & 1 & 0 \\ \alpha ab & 0 & a^2 \end{bmatrix}$$

With regard to the cardinality of the data set, each of the classes had 200 instances in the corresponding two and three-dimensional space. For the distance computations, we used the MD, which is based on the means and the covariance matrices Σ_1 and Σ_2. It is based on the correlations between the variables using which different patterns can be identified and analyzed.

In order to not make the chapter too cumbersome, the *specific* details of the values of the μ's, a, b and α (for the means and covariances), are not included here[3]. However, what is crucial to guarantee "repeatability", are the respective values of the BD for each experimental setting, and *these* are clearly specified in every single row.

Experimental Results: Artificial Data Sets. The experimental results obtained for two dimensional artificial data sets can be seen in Table 1 and those for three dimensional artificial data sets can be seen in Table 2.

Table 1. Results of the classification of two dimensional artificial data sets

BD	**1NN**	**3NN**	**SVM**	**ABBI**
4.52	100	100	100	100
2.94	99.10	99.20	99.25	99.25
1.69	95.30	96.50	97.00	96.40
0.78	84.15	86.05	88.25	88.0
0.45	73.55	75.45	81.50	80.55

Table 2. Results of the classification of three dimensional artificial data sets

Class Nature	**Average BD**	**1NN**	**3NN**	**SVM**	**ABBI**
Separated	6.08	100	100	100	100
Semi-overlapped	2.64	96.92	97.67	97.81	95.67
Overlapped	2.42	94.50	95.50	96.50	94.72
Highly overlapped	1.43	83.50	87.23	88.79	85.20

By examining Tables 1 and 2, one can see that ABBI can achieve remarkable classification when compared to that attained by the benchmark classifiers. For

[3] These values can be included if requested by the Referees.

example, if we consider the case where the classes are separated by a BD of 1.66 in Table 1, ABBI can achieve a classification accuracy of 95.38%, while the 3NN achieves 97.25%. This is quite fascinating when we consider the fact the ABBI performs the classification based *only 5 samples* from the selected 10 samples from each class, whereas the classification of NN involves the entire training set.

Real-life Data Sets: The data sets [8] used in this study have two classes, and the number of attributes varies from four up to thirty two. The data sets are described in Table 3.

Table 3. The Real-life data sets used in our experiments

Data set	No. Instances	No. Attributes	No. Classes	Attribute Type
WOBC	699	9	2	Integer
WDBC	569	32	2	Real
Diabetes	768	8	2	Integer, Real
Hepatitis	155	19	2	Categorical, Integer, Real
Iris	150	4	3	Real
Statlog (Heart)	270	13	2	Categorical, Real
Statlog (Australian Credit)	690	14	2	Categorical, Integer, Real
Vote	435	16	2	Categorical, Integer

Experimental Results: Real-life Data Sets. The results obtained for the ABBI algorithm are tabulated in Table 4.

Table 4. Classification of Real Data

Data set	kNN	NB	SVM	ABBI
WOBC	96.60	96.40	95.99	95.80
WDBC	96.66	92.97	97.71	92.39
Diabetes	75.26	73.1098	76.70	72.30
Hepatitis	82.58	83.19	82.51	80.27
Iris	95.13	96.00	96.67	94.53
Statlog (Australian Credit)	85.90	87.40	85.51	78.85
Statlog (Heart)	84.40	83.00	85.60	82.50
Vote	94.2857	90.23	94.33	90.76

From the table of results, one can see that the proposed algorithm achieves a comparable classification when compared to the other traditional classifiers, which is particularly impressive because only a very few samples are involved in the process. For example, for the WOBC data set, we can see that the new approach yielded a accuracy of 95.80% which should be compared to the accuracies of the SVM (95.99%), NB (96.40%) and the kNN (96.60%). Similarly, for the Iris data set, ABBI can achieve an accuracy of 94.53%, which is again comparable to the performance of SVM (96.67%), NB (96.00%), and NN (95.13%).

4 Conclusions

The objective of BI algorithms is to reduce the number of training vectors by selecting the patterns that are close to the class boundaries. However, the patterns

that are on the exact border of the classes ("near" borders) are not sufficient to perform a classification which is comparable to that obtained based on the centrally located patterns. In order to resolve this issue, researchers have tried to add more patterns ("far" borders) to the "border" set so as to boost the quality of the resultant border set. Thus, the cardinality of the resultant border set can be relatively high. After obtaining such a large border set, a classifier has to be generated for this set, to perform a classification.

In this paper, we have proposed a novel BI algorithm which involves the border patterns selected with respect to a new definition of the term "border". In line with the newly proposed OS-based anti-Bayesian classifiers [5,6,7], we created the "border" set by selecting those patterns which are close to the true border of the *alternate* class. The classification is achieved with regard to *these* border patterns alone, and the size of this set is very small, in some cases, as small as five from each class. The resultant accuracy is comparable to that attained by other well-established classifiers. The superiority of this method over other BI schemes is that it yields a relatively small border set, and as the classification is based on the border patterns themselves , it is computationally inexpensive.

References

1. Duch, W.: Similarity Based Methods: A General Framework for Classification, Approximation and Association. Control and Cybernetics 29(4), 937–968 (2000)
2. Foody, G.M.: Issues in Training Set Selection and Refinement for Classification by a Feedforward Neural Network. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium, pp. 409–411 (1998)
3. Foody, G.M.: The Significance of Border Training Patterns in Classification by a Feedforward Neural Network using Back Propogation Learning. International Journal of Remote Sensing 20(18), 3549–3562 (1999)
4. Li, G., Japkowicz, N., Stocki, T.J., Ungar, R.K.: Full Border Identification for Reduction of Training Sets. In: Bergler, S. (ed.) Canadian AI 2008. LNCS (LNAI), vol. 5032, pp. 203–215. Springer, Heidelberg (2008)
5. Oommen, B.J., Thomas, A.: Optimal Order Statistics-based "Anti-Bayesian" Parametric Pattern Classification for the Exponential Family. Pattern Recognition (accepted for publication, 2013)
6. Thomas, A., Oommen, B.J.: Order Statistics-based Parametric Classification for Multi-dimensional Distributions (submitted for publication, 2013)
7. 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)
8. Frank, A., Asuncion, A.: UCI Machine Learning Repository (2010), `http://archive.ics.uci.edu/ml` (April 18, 2013)
9. Devroye, L.: Non-Uniform Random Variate Generation. Springer, New York (1986)
10. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press, San Diego (1990)