Generalized Risk Zone: Selecting Observations for Classification

R.T. Peres and C.E. Pedreira, Senior Member, IEEE

Abstract—In this paper, we extend the risk zone concept by creating the Generalized Risk Zone. The Generalized Risk Zone is a model-independent scheme to select key observations in a sample set. The observations belonging to the Generalized Risk Zone have shown comparable, in some experiments even better, classification performance when compared to the use of the whole sample. The main tool that allows this extension is the Cauchy-Schwartz divergence, used as a measure of dissimilarity between probability densities. To overcome the setback concerning pdf’s estimation, we used the ideas provided by the Information Theoretic Learning, allowing the calculation to be performed on the available observations only. We used the proposed methodology with Learning Vector Quantization, feedforward Neural Networks, Support Vector Machines, and Nearest Neighbors.

Index Terms—Classification, neural networks, observations selection, risk zone, support vector machine.

1 INTRODUCTION

In this paper, we are concerned with the selection of a subset of the observations in classification environment. There are a number of previous contributions in the literature concerning observations selection under a variety of approaches and objectives, for instance, belonging to the same group of techniques, Active Learning [8], [13], [14], Query-Based Learning [9], and Sequential Design [10]. Further alternatives [21] are instance-based, lazy learning, and instance-pruning [21], [22], [23]. Notably, Support Vector Machine (SVM) [6], [7] is a machine learning tool that produces the selection of a subset of the observations, the Support Vectors (SVs).

In [1], a new method to select a subset of the training data with the goal of updating Learning Vector Quantization (LVQ) [11] prototypes was proposed with the goal of conducting the prototypes to converge at a more convenient location, diminishing misclassification errors. The selected subset is composed of observations considered to be at the risk of being captured by a prototype representing another class. This selected subset of the sample was denominated the Risk Zone (RZ). In the present contribution, we extended the RZ concept by developing what we name the Generalized Risk Zone (GRZ). The generalization comes from the fact that the GRZ is a model-free concept, not restricted to prototype methods, as is the RZ. The GRZ selects observations based on the dissimilarity between probability densities. The goal is to find the most informative observations concerning their impact on classification performance. We tested the GRZ in different models: LVQ, Layered Neural Networks, SVM, and K-NN for several data sets.

In order to do this extension, we used the Cauchy-Schwartz divergence as a measurement of the distance between two probability density functions (pdf’s), as described in Section 2. In order to avoid the usual hardship concerning the estimation of pdf’s, we used the Information Theoretic Learning (ITL) approach [2], allowing the information to be directly extracted from the available observations. The ITL is a kernel-based methodology that holds information theory concepts.

2 METHODOLOGY

Let X be a sample of size n comprised by a set of observations \{x_1, x_2, ..., x_n\} \in \mathbb{R}^m, \forall i = 1, ..., n. Let us consider a dichotomous classification environment, where each observation \(x_i\) is associated to one out of two possible classes, \(C_1\) or \(C_2\), with labels 1 or 2, respectively.

Prototype methods assign the observations using a set of prototype vectors in the observations space, in our case, the \(\mathbb{R}^m\). These prototypes are appropriately placed by using the in-sample (training) observations and are associated to one of the classes. The out-of-sample (test) observations are classified by association with the class of their nearest prototype. The RZ, as defined in [1], is a region of the space where the observations may be considered to be at the risk of being captured by the wrong class prototype. An observation is said to belong to the RZ if, and only if, its distance to a prototype of another class is smaller than the distance between this prototype and the nearest prototype of the same class [1]. Formally, let \(p_c\) be the nearest prototype representing the class of an observation \(x_i\) and \(p_r\) be any of the prototypes representing the other class. An observation \(x_i\) is said to belong to the risk zone if, and only if

\[
d(x_i, p_r) < d(p_c, p_r).
\]  

(1)

The following developments will lead to the construction of the GRZ that is not refrained to prototypes methods or any other specific model. The Cauchy-Schwartz divergence [15], defined below, allows the calculation of the distance between two pdfs \(p\) and \(q\)

\[
D_{CS}(p, q) = -\log \left( \int \frac{(p(x)q(x)dx)^2}{\int p^2(x)dx \int q^2(x)dx} \right) - \log \int p^2(x)dx - 2\log \int p(x)q(x)dx + \log \int q^2(x)dx.
\]  

(2)

Clearly, \(D_{CS}(p, p) = 0\) and \(D_{CS}(p, q) = D_{CS}(q, p)\) for any pdfs \(p\) and \(q\). Besides, \(D_{CS}(p, q) \geq 0\) [4], [15]. Nevertheless, the \(D_{CS}\) is not a metric since it does not satisfy the triangle inequality [15].

Let \(p\) and \(q\) be the pdfs associated with the observations of classes \(C_1\) and \(C_2\), respectively. It follows that the proximity between classes \(C_1\) and \(C_2\) may be measured through \(D_{CS}(p, q)\). Back to the prototypes environment, one may draw a parallel with the measurement of the distance in \(\mathbb{R}^m\) between class-representative prototypes. In the same manner, one may make an analogy for the distance between an observation and a prototype and the divergence between an observation and a pdf (associated to a class). These analogies set the basis for the definition of the GRZ, which appears farther on, in (8). Before that, we need to put forward the following developments concerning the Cauchy-Schwartz divergence.

The divergence calculation implies, in general, the necessity of estimating the involved pdfs. This is especially difficult for continuous variables due to the need for some sort of discretization procedure. The ITL approach, proposed in [2], has overcome this setback by extracting information directly from the observations. Let \(M\) and \(N\) be the sizes of the samples generated by pdfs \(p\) and \(q\), respectively. These pdfs may be estimated through a Parzen Windows approach [3]. Accordingly, let \(G_{ij} : \mathbb{R}^m \rightarrow \mathbb{R}\) be a zero mean Gaussian function with covariance matrix given by \(\sigma_i^2 I\). An estimative of the pdf \(p\), \(\hat{p}\), using \(G_{ij}\) as kernel, is given by

\[
\hat{p}(x) = \frac{1}{M} \sum_{i=1}^{M} G_{ij}(x - x_i).
\]
\[ \hat{p} = \frac{1}{M} \sum_{i=1}^{M} G_{\sigma}(x - x_i). \]

It follows that
\[
\int \hat{p}(x) \hat{q}(x) dx \\
= \int \left( \frac{1}{M} \sum_{i=1}^{M} G_{\sigma}(x - x_i) \right) \left( \frac{1}{N} \sum_{j=1}^{N} G_{\sigma}(x - x_j) \right) dx \\
= \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \int G_{\sigma}(x - x_i) G_{\sigma}(x - x_j) dx \\
= \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} G_{\sigma}(x_i - x_j). 
\]

The last equality in (3) results from the convolution theorem for Gaussians [4], [16]. It is fundamental to note that there are no approximations in the calculation of \( \int \hat{p}(x) \hat{q}(x) dx \). Furthermore, it is quite interesting to note that the final expression (3) depends exclusively on the observations \( x_i \), besides the kernel width \( \sigma \).

In a similar manner one may get
\[
\int \hat{p}^2(x) dx = \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} G_{2\sigma}(x_i - x_j) 
\]
and
\[
\int \hat{q}^2(x) dx = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} G_{2\sigma}(x_i - x_j). 
\]

By applying (3), (4), and (5) in (2), one gets
\[
D_{C-S}(\hat{p}, \hat{q}) \\
= \log \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} G_{2\sigma}(x_i - x_j) \\
- 2 \log \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} G_{\sigma}(x_i - x_j) \\
+ \log \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} G_{2\sigma}(x_i - x_j). 
\]

By applying expression (6), one may calculate the Cauchy-Schwartz divergence between two pdfs using the available observations only.

By similar arguments, one may attain the expression for the divergent between an observation \( x_k \) and an estimated pdf \( \hat{q} \):
\[
D_{C-S}(x_k, \hat{q}) = \log G_{2\sigma}(0) \\
- 2 \log \frac{1}{N} \sum_{i=1}^{N} G_{\sigma}(x_k - x_i) \\
+ \log \frac{1}{N^2} \sum_{i=1}^{N} G_{2\sigma}(x_i - x_k). 
\]

We may now define the GRZ as follows: An observation \( x_k \) is said to belong to the GRZ if, and only if,
\[
D_{C-S}(x_k, \hat{q}) < \alpha D_{C-S}(\hat{p}, \hat{q}), \quad \text{if } x_k \text{ belongs to } C_1 \quad (8.1) \\
D_{C-S}(x_k, \hat{q}) < \alpha D_{C-S}(\hat{p}, \hat{q}), \quad \text{if } x_k \text{ belongs to } C_2 \quad (8.2)
\]

Parameter \( \alpha \) in (8.1) and (8.2) was introduced for numerical reasons; note that the value of the first term of expression in (8.1) (equal argument may be used for (8.2)) may become disproportionately large in relation to the second term value. Note that we are dealing with two divergents: the first between an observation and a pdf, and the second between two pdfs. A divergent between an observation and a pdf is equivalent to a divergent between two pdfs for the case when one of these pdfs has just one observation (see expressions (6) and (7)). So, back to the expression (8.1), first term, unless a particular observation \( x_k \), say from class \( C_1 \), is very near pdf \( \hat{q} \), this term value may become extremely large. Note that the second term of the divergent in expression (7) is the negative of a logarithm of a kernel function that may assume near zero values. Concerning the divergent between two pdfs, a reasonable number of observations will generate a certain number of nonvanishing parcels in the sum (second term in (6)). For all experiments of Section 5, we used \( \alpha = 3 \).

## 3 On The Implementations

We perform a filtering process for data cleaning before GRZ calculation. This is done by applying K-NN (\( K = 3 \)) and removing all misclassified observations. This filtering is advisable since in noisy environments, small regions of the GRZ (calculated by (8.1) and (8.2)) would be created around the noisy observations. This situation, without the cleaning being applied, is illustrated in Figs. 1a and 1b, where the dark gray dots in Fig. 1a are the noisy observations (artificially inverted labels). Note that some undesirable isolated small subsets of the GRZ (in dark gray) were created, as illustrated in Fig. 1b. This setback was overcome by eliminating the noisy observations (see Fig. 1c).

Four classification methods were implemented (find results in Section 5) using both: the observations that belong to the GRZ only and the whole sample set. The four implemented methods are:

1. **LVQ:**
2. a feedforward neural network (NN), trained with Bayesian regularization, with 10 initial neurons in the hidden layer
Concerning the sigma width, we used four approaches, the first two based on the approach proposed in [20]:

1. cross validation;
2. optimal window width assuming that data are normally distributed with unit variance and that the kernel is Normal ((20, (4.14)));
3. averaged value of 1 and 2; and
4. fixed value (0.1).

For each experiment, we used the sigma widths taking in account a trade-off between performance and the number of selected observations.

For experiments 1-5, we used, for LVQ and Neural Networks, the average value over the ten independent initializations. For SVM, we varied the parameters value (kernel width and constant C) for the following values: $2^k$, $2^{k+3}$, $2^{k+2}$, $2^k$, $2^{k-1}$ and $2^{k-2}$ and took the out of sample performance that corresponded to the highest performance in sample. K-NN was run with $K = 3$.

For experiments 6-8 (the ones with fewer observations), we used a 10-fold cross validation. Accordingly, for each data set, a 10-fold disjoint partition of the data is created and, for each technique (LVQ, NN, SVM, and K-NN), the in-sample phase was performed by a training set using 9 of the 10 folds of the partition. The remaining fold was used to calculate the out-of-sample performance. This procedure was repeated 10 times, each of those with a different fold as test set. For all the techniques, the average accuracy is reported. Concerning SVM, kernel width and constant C were both varied for $2^k$, $2^{k+3}$, $2^{k+2}$, $2^k$, $2^{k-1}$ for each of the 10-fold simulations. The best parameters combination for the in-sample performance was used to calculate the out-of-sample performance. Finally, for experiment 9, we used a classic leave-one-out procedure [3] with the 27 observations.

4 On the Data Sets
In the next section, we present the numerical results for nine data sets: two controlled experiments, plus seven public available data sets from the UCI repository: Waveform, Letter-B, Statlog, Heart diseases diagnosis, Breast Cancer, Ionosphere, and Lung Cancer.

The data set for experiment 1 consists of two classes divided by a cosine function. We generated 1,030 observations of class C1 and 1,027 observations of class C2 for in-sample, and 1,060 observations labeled C1 and 1,041 observations labeled C2 for out-of-sample. For experiment 2, two classes were generated through a circle and a roll with coincident centers (without superposition). In-sample: 123 observations of class C1 and 2,611 observations of class C2. Out-of-sample: 127 observations of class C1 and 2,646 observations of class C2.

Experiment 3 consists on waveform data set [25]. There are three classes of waves (here, we used class 1 and the others as class 2), 5,000 observations (3,000 for in-sample phase and 2,000 for the out-of-sample phase), and 40 input features, all of which include noise. The latter 19 input features are all noise attributes with mean 0 and variance 1.

Experiment 4 is related to letter recognition. The objective is to identify each of a large number of black and white rectangular pixel displays as one of the 26 capital letters in the English alphabet. Since here we are concerned with a binary environment, we choose letter B to represent the class C1 and all of the other 25 letters are C2. The data set consists of 20,000 observations


The data set for experiment 5 is the Statlog—landsat satellite. This data set consists of the multispectral values of pixels in $3 \times 3$ neighborhoods in a satellite image, and the classification associated with the central pixel in each neighborhood. The aim is to predict this classification, given the multispectral values. Once again, since we are concerned with a two classes problem, the original six classes were reduced to two, class 1 against the others. There are 6,435 observations (4,435 for in-sample phase and 2,000 for the out-of-sample test) and 36 input features.

The data set for experiment 6 was formed by assembling four data sets [12] concerning heart diseases diagnosis. Each of these four data sets is individually available in the UCI machine learning repository. The data were collected from the Cleveland Clinic Foundation, the Hungarian Institute of Cardiology, the V.A. Medical Center, and the Zurich University Hospital. All of the original databases have 76 features, but only 13 of them are actually relevant [5]. The goal is to predict angiographic disease status concerning narrowing in major vessels. After missing data elimination, the data set ended up with 740 patients and 10 input features.

The experiment 7 data set provided by the University of Wisconsin Hospitals, Madison, has been frequently used as a benchmark [1], [17], [18], [19]. Data are related to an application in diagnosing breast mass cytology, the features are nine cytological characteristics of benign or malignant breast fine-needle aspirates (uniformity of cell shape, uniformity of cell size, lump thickness, bare nuclei, cell size, normal nucleioli, lump cohesiveness, nuclear chromatin, and mitosis). All of the features assume discrete values between 1 and 10. The two output classes are "benign" or "malignant" tumor. After removing the observations with missing values, the data set end up with 683 observations.

The data set for experiment 8 refers to data collected by a radar system in Goose Bay, Labrador. This system consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kilowatts. The targets were free electrons in the ionosphere. "Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not; their signals pass through the ionosphere. Received signals were processed using an autocorrelation function whose arguments are the time of a pulse and the pulse number. There were 17 pulse numbers with observations described by two features per pulse number resulting in 34 input features. There are 351 observations.

The Lung Cancer data set (experiment 9) was used in [24] to illustrate the power of the optimal discriminant plane even in ill-posed settings. The data describe three types of pathological lung cancers using 56 input features. Here, we used pathology 1 against the others in the binary classification problem. The data set has originally 32 observations, resulting in 27, 8 from C1, and 19 from C2, after removing the missing values.

5 Results
Performance results, for all data sets and all the used methods (LVQ, Layered Neural Networks, SVM, and K-NN) may be found in Tables 1, 2, 3, 4. These tables include in-sample and out-of-sample correctness rates, as well as the standard deviation concerning the k-fold experiments. Results are presented using: 1) all available observations (column "Without GRZ" in Tables 1, 2, 3, 4) and 2) observations belonging to the GRZ without filtering (column "With GRZ") and the observations belonging to the GRZ with filtering (column "With GRZ + Filtering"). We also show a comparison with the IB3 [22] and DROP3 [23] methods (Table 5). Both algorithms have a good trade-off between generalization accuracy and observation selection.

### TABLE 1
LVQ Experiments Out-of-Sample Correctness Rate (Parenthesis Indicates In-Sample)

<table>
<thead>
<tr>
<th></th>
<th>Without GRZ</th>
<th>With GRZ</th>
<th>With GRZ + Filtering</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average Correctness</td>
<td>std</td>
<td>Percentual of used observations</td>
</tr>
<tr>
<td>Exp 1 Cosine</td>
<td>97.8% (98.4%)</td>
<td>0.3</td>
<td>100%</td>
</tr>
<tr>
<td>Exp 2 Circle</td>
<td>97.9% (98.1%)</td>
<td>0.1</td>
<td>100%</td>
</tr>
<tr>
<td>Exp 3 Waveform</td>
<td>91.5% (93.0%)</td>
<td>0.3</td>
<td>100%</td>
</tr>
<tr>
<td>Exp 4 Letter</td>
<td>96.6% (98.1%)</td>
<td>0.1</td>
<td>100%</td>
</tr>
<tr>
<td>Exp 5 Statlog</td>
<td>98.2% (98.4%)</td>
<td>0.2</td>
<td>100%</td>
</tr>
<tr>
<td>Exp 6 Heart</td>
<td>78.0% (83.6%)</td>
<td>4.6</td>
<td>100%</td>
</tr>
<tr>
<td>Exp 7 Breast Cancer</td>
<td>97.1% (97.3%)</td>
<td>2.0</td>
<td>100%</td>
</tr>
<tr>
<td>Exp 8 Ionosphere</td>
<td>86.3% (89.7%)</td>
<td>5.1</td>
<td>100%</td>
</tr>
<tr>
<td>Exp 9 Lung Cancer</td>
<td>77.8% (96.0%)</td>
<td>-</td>
<td>100%</td>
</tr>
</tbody>
</table>

### TABLE 2
NN Experiments Out-of-Sample Correctness Rate (Parenthesis Indicates In-Sample)

<table>
<thead>
<tr>
<th></th>
<th>Without GRZ</th>
<th>With GRZ</th>
<th>With GRZ + Filtering</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average Correctness</td>
<td>std</td>
<td>Percentual of used observations</td>
</tr>
<tr>
<td>Exp 1 Cosine</td>
<td>99.4% (99.7%)</td>
<td>0.3</td>
<td>100%</td>
</tr>
<tr>
<td>Exp 2 Circle</td>
<td>99.7% (100%)</td>
<td>0.1</td>
<td>100%</td>
</tr>
<tr>
<td>Exp 3 Waveform</td>
<td>89.7% (99.6%)</td>
<td>0.5</td>
<td>100%</td>
</tr>
<tr>
<td>Exp 4 Letter</td>
<td>98.6% (99.8%)</td>
<td>0.2</td>
<td>100%</td>
</tr>
<tr>
<td>Exp 5 Statlog</td>
<td>98.8% (99.9%)</td>
<td>0.3</td>
<td>100%</td>
</tr>
<tr>
<td>Exp 6 Heart</td>
<td>75.3% (97.1%)</td>
<td>5.1</td>
<td>100%</td>
</tr>
<tr>
<td>Exp 7 Breast Cancer</td>
<td>94.7% (99.6%)</td>
<td>3.0</td>
<td>100%</td>
</tr>
<tr>
<td>Exp 8 Ionosphere</td>
<td>86.9% (100%)</td>
<td>4.7</td>
<td>100%</td>
</tr>
<tr>
<td>Exp 9 Lung Cancer</td>
<td>85.2% (100%)</td>
<td>-</td>
<td>100%</td>
</tr>
</tbody>
</table>
### TABLE 3

<table>
<thead>
<tr>
<th></th>
<th>Without GRZ</th>
<th></th>
<th>With GRZ</th>
<th></th>
<th>With GRZ + Filtering</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average Correctness</td>
<td>std</td>
<td>Percentual of used observations</td>
<td>Average Correctness</td>
<td>std</td>
<td>Percentual of used observations</td>
</tr>
<tr>
<td>Exp 1 Cosine</td>
<td>99.0% (99.2%)</td>
<td>-</td>
<td>100%</td>
<td>84.4% (97.7%)</td>
<td>-</td>
<td>42.7%</td>
</tr>
<tr>
<td>Exp 2 Circle</td>
<td>99.0% (98.7%)</td>
<td>-</td>
<td>100%</td>
<td>99.9% (98.8%)</td>
<td>-</td>
<td>9.0%</td>
</tr>
<tr>
<td>Exp 3 Waveform</td>
<td>67.6% (100%)</td>
<td>-</td>
<td>100%</td>
<td>67.6% (100%)</td>
<td>-</td>
<td>85.9%</td>
</tr>
<tr>
<td>Exp 4 Letter</td>
<td>99.4% (100%)</td>
<td>-</td>
<td>100%</td>
<td>98.9% (100%)</td>
<td>-</td>
<td>27.5%</td>
</tr>
<tr>
<td>Exp 5 Statlog</td>
<td>98.3% (100%)</td>
<td>-</td>
<td>100%</td>
<td>53% (100%)</td>
<td>-</td>
<td>20.6%</td>
</tr>
<tr>
<td>Exp 6 Heart</td>
<td>75.0% (100%)</td>
<td>5.1</td>
<td>100%</td>
<td>70.1% (100%)</td>
<td>4.8</td>
<td>42.5%</td>
</tr>
<tr>
<td>Exp 7 Breast Cancer</td>
<td>95.9% (100%)</td>
<td>3.1</td>
<td>100%</td>
<td>95.3% (100%)</td>
<td>3.8</td>
<td>66.6%</td>
</tr>
<tr>
<td>Exp 8 Ionosphere</td>
<td>89.2% (100%)</td>
<td>7.7</td>
<td>100%</td>
<td>66.4% (100%)</td>
<td>6.9</td>
<td>63.3%</td>
</tr>
<tr>
<td>Exp 9 Lung Cancer</td>
<td>70.4% (100%)</td>
<td>-</td>
<td>100%</td>
<td>70.4% (100%)</td>
<td>-</td>
<td>58.8%</td>
</tr>
</tbody>
</table>

### TABLE 4

<table>
<thead>
<tr>
<th></th>
<th>Without GRZ</th>
<th></th>
<th>With GRZ</th>
<th></th>
<th>With GRZ + Filtering</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average Correctness</td>
<td>std</td>
<td>Percentual of used observations</td>
<td>Average Correctness</td>
<td>std</td>
<td>Percentual of used observations</td>
</tr>
<tr>
<td>Exp 1 Cosine</td>
<td>99.0% (99.0%)</td>
<td>-</td>
<td>100%</td>
<td>99.0% (93.9%)</td>
<td>-</td>
<td>16.7%</td>
</tr>
<tr>
<td>Exp 2 Circle</td>
<td>99.5% (99.3%)</td>
<td>-</td>
<td>100%</td>
<td>99.5% (92.7%)</td>
<td>-</td>
<td>9.0%</td>
</tr>
<tr>
<td>Exp 3 Waveform</td>
<td>88.0% (86.0%)</td>
<td>-</td>
<td>100%</td>
<td>88.0% (83.7%)</td>
<td>-</td>
<td>85.9%</td>
</tr>
<tr>
<td>Exp 4 Letter</td>
<td>99.3% (99.3%)</td>
<td>-</td>
<td>100%</td>
<td>99.2% (97.4%)</td>
<td>-</td>
<td>27.5%</td>
</tr>
<tr>
<td>Exp 5 Statlog</td>
<td>99.6% (98.9%)</td>
<td>-</td>
<td>100%</td>
<td>98.6% (94.4%)</td>
<td>-</td>
<td>20.6%</td>
</tr>
<tr>
<td>Exp 6 Heart</td>
<td>76.5% (76.4%)</td>
<td>3.4</td>
<td>100%</td>
<td>74.3% (60.6%)</td>
<td>4.4</td>
<td>42.5%</td>
</tr>
<tr>
<td>Exp 7 Breast Cancer</td>
<td>96.6% (96.9%)</td>
<td>2.5</td>
<td>100%</td>
<td>96.5% (95.2%)</td>
<td>2.7</td>
<td>66.6%</td>
</tr>
<tr>
<td>Exp 8 Ionosphere</td>
<td>83.8% (84.7%)</td>
<td>5.7</td>
<td>100%</td>
<td>82.1% (87.8%)</td>
<td>5.0</td>
<td>59.1%</td>
</tr>
<tr>
<td>Exp 9 Lung Cancer</td>
<td>85.2% (84.6%)</td>
<td>-</td>
<td>100%</td>
<td>85.2% (83.4%)</td>
<td>-</td>
<td>93.0%</td>
</tr>
</tbody>
</table>
6 DISCUSSION AND FINAL REMARKS

The GRZ methodology has shown very good results when applied to some widely used techniques. The performance of GRZ + Filtering is, in general, slightly better with a considerable reduction in the number of used observations. In general, the GRZ without Filtering is not as good as GRZ plus Filtering, although for some experiments, the GRZ without filtration produced quite good results (see, for experiment 9, Table 2 and Table 3 for experiment 7). Concerning LVQ, the proposed GRZ + Filtering methodology has provided encouraging results with a substantial reduction in the number of used observations. For instance, in experiment 2 in Table 1, a better performance was reached using only 9.1 percent of the observations. Also, for NN (Table 2) and K-NN (Table 4), the GRZ has provided a substantial reduction in the number of observations with comparable (better for some experiments) performance. In this sense, the proposed methodology was able to identify appropriated subsets of the sample in which the performance did not drop. For SVM, the best relative performance was obtained for experiment 6 (see Table 3), an almost 4 percent gain in the correctness rate with just half of the observations. In comparison with the GRZ, the IB3 and DROP3 are completely different approaches from the theoretical point of view. Experimentally, IB3 and DROP3 have shown, in general, a significant reduction in the number of observations still maintaining good performance. The GRZ + Filtering also provided important reductions in the number of observations, as shown in Tables 1, 2, 3, 4, but not as much as IB3 and DROP3. Nevertheless, concerning correctness, the GRZ + Filtering has slightly beaten the IB3 in 27 out of 36 experiments (two methods times nine experiments) and the DROP3 in 21 out of 9. In this situation, the GRZ + Filtering beat the IB3 in eight out of nine experiments and the DROP3 in six out of nine. In this paper, we proposed and implemented, in a classification environment, a methodology to identify representative observations in a data set. This methodology extends the Risk Zone concepts, proposed in [1], allowing its use in approaches other than prototype models. The key tool to reach this goal was the use of the divergence between two pdfs, instead of metrical distances. The setback concerning pdfs estimation was overcome by employing Information Theoretic Learning concepts, allowing the calculation to be restricted to the available observations.

ACKNOWLEDGMENTS

The work of C.E. Pedreira was partially supported by grants from the Brazilian National Research Council (CNPq) and the Rio de Janeiro Research Foundation (FAPERJ). The work of R.T. Peres was supported by a PhD scholarship from CNPq. The authors are also in debt to Professor Alexandre Pinto Alves da Silva and to Mr. Helio Pinto for allowing some of the experiments performed in their Lab (LASPOT-UFRJ).

REFERENCES


⇒ For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.