THE GENERALIZED RISK ZONE AND OBSERVATIONS SELECTION

RODRIGO T. PERES[(1)] AND CARLOS E. PEDREIRA[(2)]

(1) Departamento de Engenharia Elétrica, PUC-Rio
(2) Faculdade de Medicina e COPPE-PEE, Universidade Federal do Rio de Janeiro
E-mail: pedreira@ufrj.br

Abstract – We extend the risk zone concept by creating the Generalized Risk Zone. The Generalized Risk Zone is applied in a model-independent methodology to select representative observations in a sample set, with the goal of enhancing classification performance. The methodology involves the calculation of Cauchy-Schwartz divergence, as a measure of distance between densities, and we do so in the context of Information Theoretic Learning. We applied this methodology in Neural Networks, Support Vector Machines and Learning Vector Quantization. We have also discussed the comparison between Support Vectors and the vector that lay in the Generalized Risk Zone.

Keywords – Risk Zone, Neural Networks, Support Vector Machine, Classification, Observations Selection.

Resumo – Nós estendemos o conceito de zona de risco através da Zona de Risco Generalizada. A Zona de Risco Generalizada é aplicada em uma metodologia independente de modelo para selecionar observações representativas em uma amostra, com o objetivo de melhorar a performance de classificação. A metodologia envolve o cálculo da divergência de Cauchy-Schwartz como uma medida de distância entre densidades, implementada dentro do contexto de *Information Theoretic Learning.* Nós aplicamos esta metodologia em Redes Neurais, Máquina de Vetor de Suporte e Aprendizado por Quantização Vetorial. Discutimos também a comparação entre vetores de suporte e o vetor gerado pela Zona de Risco Generalizada.

Palavras-chave – Zona de Risco, Redes Neurais, Máquina de Vetor de Suporte, Classificação, Seleção de Observações.

1 Introduction

There may be several reasons to select subsets of observations in a sample. For instance, one may be interesting in splitting representative from non-representative observations with the goal of enhancing classification performance.

There are a number of previous contributions in the literature concerning data selection under a variety of approaches, e.g. *Active Learning* [13], [18], [19]; *query-based learning* [14] and *sequential design* [15]. It is worth noting that *Support Vector Machines* (SVM) [11], [12] implicitly select representative observations, the support vectors.

Risk zone, an idea originally proposed in [1] in a Learning Vector Quantization (LVQ) [16] context, is a key concept for the methodology proposed in this paper. Risk zone is connected to the selection of a subset of the observations with the goal of conducting the location of prototypes to convenient locations other than the class mean. This methodology was successfully applied in a heart diseases diagnosis problem [17]. Some central tools used in the sequence are in the context of *Information Theoretic Learning* (ITL) [4], [5], [7]-[9]. ITL is a kernel based methodology that lays hold of information theory concepts.

In this paper we extend the risk zone ideas by developing what we call '*Generalized Risk Zone'* (GRZ). GRZ is applied in a model-independent methodology to select representative observations in a sample set.

2 Methodology

The Cauchy-Schwartz divergence [9] is a fundamental tool in the GRZ development. It allows the calculation of 'distances' between different probability density functions (pdfs). In fact, the Cauchy-Schwartz divergence is not a metric distance since it does not satisfy the triangle inequality.

Let $x_i \in R^n$ be an observation and consider a supervised classification environment, with a dichotomous labeled set:

$$X = \{(x_i, y_i),\ i = 1,\ldots,N\} \qquad (1)$$

where $y_i = 1$ or $y_i = 2$. In a LVQ context one is interested in updating prototypes location in order to force these vectors to represent different classes, two of those in a dichotomous labeled sample. A *risk zone* is a region of the space where observations may be considered to be at the risk of being captured by the wrong class prototype [1]. The idea is to update the prototypes by using only the subset of the observations that belongs to the risk zone. We say that an observation x_i belongs to the risk zone if, and only if, its distance to a prototype of other class is smaller than the distance between this prototype and the nearest prototype of the same class. 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 belongs to the risk zone if, and only if $d(x_i, p_r) < d(p_c, p_r)$ for any prototype $p_r \neq p_c$. Equivalently,

$$\frac{d(x_i, p_r)}{d(p_c, p_r)} < 1.$$

Up to this point we followed [1] and so we are circumscribed to prototypes models. We follow by proposing the *Generalized Risk Zone* (GRZ) that is not restricted to this type of models.

2.1 The Generalized Risk Zone

Let us consider a supervised classification dichotomous problem as in equation (1). Call p the pdf of the observations in class 1, and q the pdf referred to class 2. The starting point is to consider the Cauchy Schwartz divergence[♣] as a measure of distance between densities instead of considering the distance between prototypes. In analogy to distance between an observation and a prototype (representative of a class), we propose the divergence between an observation and a pdf (associated to a class). We denote the Cauchy Schwartz divergence as $D_{C\text{-}S}(\bullet,\bullet)$.

We may now define the *Generalized Risk Zone* as follows: Say that an observation x_i belongs to the GRZ if, and only if

$$\frac{D_{C\text{-}S}(x_i, h)}{\left(D_{C\text{-}S}(p,q)\right)^2} < 1. \qquad (2)$$

Here, we set $h = q$ if $x_i \in$ class 1 and $h = p$ if $x_i \in$ class 2.

The square in the denominator of (2) was introduced by technical reasons related to the Cauchy Schwartz divergence calculation.

We calculate the Cauchy Schwartz divergence, in order to determine the GRZ, by using the ITL approach which involves a kernel function. The only free parameter is this kernel function width, we call σ^2 (see the appendix for details).

2.2 Concerning the implementations

We implemented three different classification algorithms: (i) LVQ 1, with three prototypes per class, 100 epochs and learning rate 0.01; (ii) A neural network (NN), trained with Bayesian regularization, with 10 initial neurons in the hidden layer and logsig activation function in the hidden and output layers and (iii) Support Vector Machine (SVM) with radial basis function kernel. We implemented k-fold experiments. All percentages are mean of 10 times (in LVQ 1 case, 10 random initialization of the prototypes choice), except for SVM, that was the best of six experiments varying C (= 0.1; = 5) and kernel size (= 0.5; = 1; = 5). Each of the methods (LVQ, NN and SVM) was implemented using just the observations that belong to the GRZ, and with the goal of comparing performances, all the methods were also implemented with all available observations.

[♣] Please find details on the Cauchy Schwartz divergence calculation in the appendix.

3 Results and Discussion

It follows the results concerning two controlled and two real data experiments.

Experiment 1: We generated two classes divided by a cosine function (in-sample: 1030 observations C_1 and 1027 observations C_2, out-of-sample: 1060 observations C_1 and 1041 observations C_2). In Figures 1a, 1b and 1c one can find the GRZ (in purple) for different the kernel sizes: σ^2 = 0.0025, 0.01 and 0.06. Classification with the algorithms trained by all observations and trained by GRZ observations can be found in table 1.

Table 1: Out-of-sample accuracy percentage (parenthesis indicates in-sample).

	All Obs 2057	**σ^2 = 0.0025** 449 obs (21.8%)[♠]	**σ^2= 0.01** 702 obs (34.1%)[♠]	**σ^2 = 0.06** 1196 obs (58.1%)[♠]
LVQ 1	92.6 (92.5)	90.9 (70.3)	91.4 (77.5)	91.8 (85.6)
NN	99.4 (99.9)	98.9 (98.2)	99.2 (99.1)	99.4 (99.6)
SVM	99.5 (99.6)	99.5 (98.2)	99.5 (98.9)	99.5 (99.3)

♠ Of the total number of observations (2057)

Notably SVM has shown equivalent out-of-sample performance for training with all observations and with just the GRZ selected (for different values of σ^2). SVM algorithms are time consuming and this approach may represent a significant gain from the computational point of view. One hundred and thirty support vectors resulted from the run with for all observations. Conversely, 101 (77.7%) of the support vectors were selected to be at the GRZ for all values of σ^2. When SVM was run only using GRZ observations, we had 100% of coincidence between those and the support vectors for σ^2 = 0.0025 and σ^2 = 0.01 and 99% for σ^2 = 0.06.

Another point to be noted is that, by applying the GRZ selection, equivalent performance was obtained using just a fraction of the observations. This shows that the observations in the GRZ are really the relevant ones.

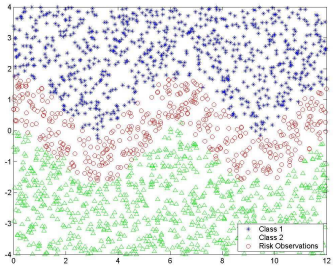

Fig. 1a: Classes 1 and 2 and GRZ for σ^2 = 0.0025 (449 observations).

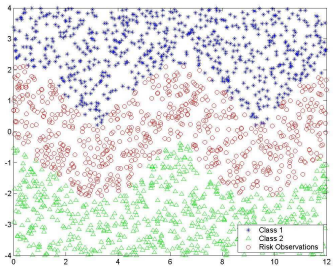

Fig. 1b: Classes 1 and 2 and GRZ for σ^2 = 0.01 (702 observations).

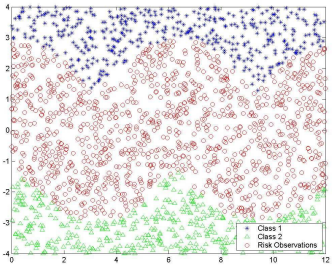

Fig. 1c: Classes 1 and 2 and GRZ for σ^2 = 0.06 (1196 observations).

Experiment 2: We generated two classes by using a circle and a roll with the same centers and no superposition. For in-sample phase we had 123 observations C_1 and 2611 observations C_2. For out-of-sample phase were 127 observations C_1 and 2646 observations C_2. In Figures 2a to 2c one can find the GRZ for σ^2 = 0.0025, 0.01 and 0.06. Classification with the algorithms trained by all observations and trained by GRZ observations can be found in table 2.

Table 2: Out-of-sample accuracy percentage (parenthesis indicates in-sample).

	All Obs 2734	**σ^2 = 0.0025** 280 obs (10.2%)♠	**σ^2 = 0.01** 453 obs (16.6%)♠	**σ^2 = 0.06** 784 obs (28.7%)♠
LVQ 1	88.1 (87.5)	98.2 (86.8)	97.8 (89.8)	97.4 (93)
NN	99.7 (100)	99.3 (100)	99.6 (100)	99.6 (100)
SVM	99.6 (99.7)	99.6 (97.1)	99.6 (98.5)	99.6 (99)

♠ Of the total number of observations (2734)

Note that, training with the GRZ subsets has clearly enhanced the LVQ performance and matched the NN and SVM ones. Three hundred and ten support vectors resulted from the run with for all observations and 49 (15.8%) of the support vectors were selected to be at the GRZ for σ^2 = 0.0025 and σ^2 = 0.01. For σ^2 = 0.06, 96 (31%) of the support vectors were in GRZ. When SVM was run only using GRZ observations, we had 100% of coincidence between those and the support vectors for all observations for σ^2 equal to 0.0025, 83.1% for σ^2 equal to 0.01 and 63.4% for σ^2 equal to 0.06.

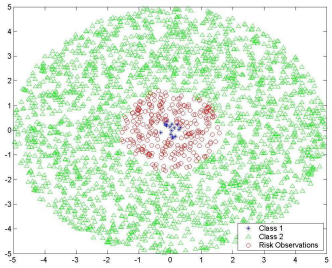

Fig. 2a: Classes 1 and 2 and GRZ for σ^2 = 0.0025 (280 observations).

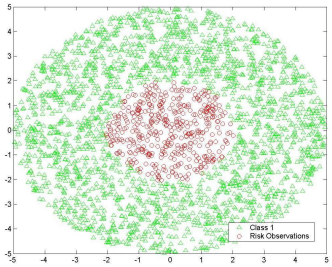

Fig. 2b: Classes 1 and 2 and GRZ for σ^2 = 0.01 (453 observations).

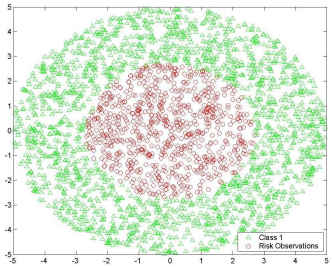

Fig. 2c: Classes 1 and 2 and GRZ for σ^2 = 0.06 (784 observations).

Experiment 3: Heart Disease Diagnosis Data Set

This data set was formed by assembling four data sets [17] concerning heart diseases diagnosis. Each of these four data sets is individually available in the UCI♣ machine learning repository. The data was collected from the Cleveland Clinic Foundation; the Hungarian Institute of Cardiology; the V.A. Medical Center, and the Zurich University Hospital. All the

♣ http://www.ics.uci.edu/~mlearn/MLSummary.html.

original databases have 76 attributes but only 13 of them are actually relevant [10]. The goal is to predict angiographic disease status concerning narrowing in major vessels. After missing data elimination, we ended up, in the assembled data set, with 740 patients and 10 input attributes.

For in-sample phase we set apart 540 patients and 200 for out-of-sample phase. Results are presented in table 3.

Table 3: Out-of-sample accuracy percentage (parenthesis indicates in-sample).

	All Obs 540	**σ^2 = 0.0025** 480 obs (88.9%)♠	**σ^2 = 0.01** 522 obs (96.7%)♠	**σ^2 = 0.006** 486 obs (90%)♠
LVQ 1	80 (80.7)	80.4 (79.4)	79.7 (80.7)	80.9 (79.6)
NN	74.6 (97.9)	71.9 (98)	74.6 (98.2)	71.5 (98.2)
SVM	81 (80.7)	82.5 (78.1)	81.5 (80.3)	82.5 (78.8)

♠ Of the total number of observations (540)

For all three classification schemes, the out-of-sample results with a reduced set of observations (by using the GRZ) are comparable, or marginally better then the results with the totality of the observations. Three hundred and ninety seven support vectors resulted from the run with for all observations and 359 (90.4%), 383 (96.5%) and 364 (91.7%) of the support vectors were selected to be at the GRZ for σ^2 equal to 0.0025, 0.01 and 0.06. When SVM was run only using GRZ observations, we had 96.2% of coincidence between those and the support vectors for all observations for σ^2 equal to 0.0025, 98.7% for σ^2 equal to 0.01 and 97% for σ^2 equal to 0.06.

Experiment 4: Breast Cancer Data Set

This data set concerns an application in diagnosing breast mass cytology. Data is available at the UCI♣ repository and was provided by the University of Wisconsin Hospitals, Madison. The input is composed by nine cytological characteristics of benign or of malignant breast fine-needle aspirates. These attributes are: uniformity of cell shape, uniformity of cell size, clump thickness, bare nuclei, cell size, normal nucleoli, clump cohesiveness, nuclear chromatin, and mitosis. All the attributes assume discrete values between 1 and 10. The two output classes are “benign” or “malignant”.

After removing missing points, the data set remained with 683 instances with 9 input attributes. For in-sample phase we set apart 558 patients and 125 for out-of-sample phase. Results are in table 4.

Table 4: Out-of-sample accuracy percentage (parenthesis indicates in-sample).

	All Obs 558	**σ^2 = 0.01** 243 obs (43.5%)♠	**σ^2 = 0.06** 554 obs (99.3%)♠
LVQ 1	95.3 (96.6)	96.1 (94.4)	95.4 (96.9)
NN	94.9 (99.7)	87.7 (99.6)	94.6 (99.6)
SVM	96 (97.7)	96 (93.8)	96 (97.7)

♠ Of the total number of observations (558)

Once again, a part from a slightly worst result for the NN model with σ^2 = 0.01, it seems that one may obtain comparable results by using an appropriate subset of observations. One hundred sixty eight support vectors resulted from the run with for all observations and 55 (32.7%) and 164 (97.6%) of these support vectors were selected to be at the GRZ for σ^2 equal to 0.01 and 0.06. When SVM was run only using GRZ observations, we had 69.3% of coincidence between those and the support vectors for all observations for σ^2 equal to 0.01 and 99.4% for σ^2 equal to 0.06.

4 Final Remarks

In this paper we propose a methodology to identify representative observations from a data set. It is a model-independent risk zone approach, that we call ‘Generalized Risk Zone’. The objective is to select observations that are on the risk to be wrongly classified and to investigate how useful they are in the training process.

Since the method is model-independent, GRZ observations can be used for any classification model. Experiments with LVQ, Neural Networks and SVM have been done and our results show that using GRZ observations for training at least equalize (and sometimes improve it) the accuracy for training with all observations.

A comparison with support vectors selected by SVM is also done. Since we are dealing with observations that are important for classification, we investigate both subsets from data and come up with interesting conclusions. At first we observed that in most experiments, a great percent of the support vectors selected by SVM from all observations were in GRZ observations. Besides, we investigate new sets of support vectors when SVM was directly applied to GRZ observations and compare the new support vectors with the ones obtained from all observations. It is interesting that sometimes the accuracy of classification can be kept with a smaller set of original support vectors or with a merge set composed by original SV and others observations.

♣ http://www.ics.uci.edu/~mlearn/MLSummary.html.

Future works are closely related to establish margin concepts for GRZ and to extend it to multiples classes problems.

References

[1] Pedreira, C. E. (2006) Learning vector quantization with training data selection, *IEEE Trans. Pattern Analysis and Machine Intelligence* (January), vol. 28, issue 1, pp. 157-162.

[2] Cover, T.M. e Thomas J.A. (1991) Elements of Information Theory, Wiley Series in Telecommunications.

[3] Haykin, S. (1998) Neural Networks: A Comprehensive Foundation, Prentice-Hall.

[4] Principe, J. C., Xu, D. e Fisher, J. (2000) Information Theoretic Learning, in *Unsupervised Adaptive Filtering*, S. Haykin, Ed. New York: Wiley.

[5] Morejon R. A. e Príncipe, J. C. (2004) Advanced Search Algorithms for Information Theoretic Learning with Kernel-Based Estimators, *IEEE Transactions on Neural Networks*, 15(4) pp. 874–884.

[6] Duda, R.O., Hart, P.E., Stork, G. (2001) Pattern Recognition. 2nd. Ed., Wiley.

[7] Jensen, R. (2005) An Information Theoretic Approach to Machine Learning, Dissertation for the Degree of Doctor Scientiarum.

[8] Gokcay, E., Príncipe, J. C. (2002) Information Theoretic Clustering, *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol.24, no. 2, pp. 158-171, February.

[9] Lehn-Schioler, T., Hedge, A., Erdogmus D. e Principe, J. C. (2005) Vector Quantization using Information Theoretic Concepts, *Natural Computing*, nº4, pp.39-51.

[10] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S. e Froelicher, V. (1989) International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease, *Am. J. Cardiology*, pp. 304-310, 1989.

[11] Vapnik, V. N. (1998) Statistical Learning Theory, New York: Wiley, 1998.

[12] Burges, C. J. C. (1998) A tutorial on support vector machines for pattern recognition, *Knowledge Discovery and Data Mining*, vol. 2, no. 2, pp.121–167.

[13] Plutowski, M. e White, H. (1993) Selecting concise training sets from clean data, *IEEE Trans. Neural Networks*, vol. 4, issue 2, pp. 305-318, March.

[14] Hwang, J. N., Choi, J. J., Oh, S. e Marks II, R. J. (1991) Query-based learning applied to partially trained multi-layer perceptrons, *IEEE Trans. Neural Networks*, vol.2, issue 1, pp.131-136, January.

[15] Faraway, J. J. (1990) Sequential design for the nonparametric regression of curves and surfaces, in *Proceedings of the 22nd Symposium on the Interface between Computing Science and Statistics, Springer*, pp. 104-110.

[16] Kohonen, T., (2001) Self-Organizing Maps, third ed. Springer.

[17] Pedreira, C. E., Macrini, L., Costa, E. S. (2005) Input and data selection applied to heart disease diagnosis. *Proc. IEEE-INNS-ENNS International Joint Conference on Neural Networks*, Montreal.

[18] Mitra, P., Pal S.K. (2004) A probabilistic active support vector learning algorithm, *IEEE Trans. Pattern Analysis and Machine Intelligence* (March), vol. 26, issue 3, pp. 413-418.

[19] Li, M., Sethi, I. K. (2006) Confidence-Based Active Learning, *IEEE Trans. Pattern Analysis and Machine Intelligence* (August), vol. 28, issue 8, pp. 1251-1261.

Appendix – Divergence calculation

Different measures of information, e.g., entropy, mutual information and divergence [2], [3], involve the necessity of pdf's estimation. This is especially harassing for continuous variables, since the calculation involves some sort of discretization procedure. The ITL approach proposed in [4] has overcome this setback by extracting information directly from the observations.

Let us consider the Cauchy-Schwartz divergence, between two pdfs p and q, as defined in [9]:

$$D_{C\text{-}S}(p,q) = -\log\left(\frac{\left(\int p(x)q(x)dx\right)^2}{\int p^2(x)dx \int q^2(x)dx}\right) \qquad (3)$$

Note that the $D_{C\text{-}S}$, is a way of measuring the distance between the pdf's p and q.

Let M and N be the number of available observations generated by pdf's p and q respectively. In order to estimate p and q in (3), we use a Parzen windows approach [6] with a zero mean normal kernel function. In this way one may write the estimative

$$\hat{p} = \frac{1}{M}\sum_{i=1}^{M} G(x - x_i, \sigma^2)$$

where G is a zero mean normal function. It follows

$$\int p(x)q(x)dx \approx \int \hat{p}(x)\hat{q}(x)dx =$$

$$\int\left(\frac{1}{M}\sum_{i=1}^{M} G(x - x_i, \sigma^2)\right)\left(\frac{1}{N}\sum_{j=1}^{N} G(x - x_j, \sigma^2)\right)dx$$

$$= \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\int G(x - x_i, \sigma^2)G(x - x_j, \sigma^2)dx$$

$$= \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} G(x_i - x_j, 2\sigma^2)\cdot \qquad (4)$$

The last equality results from the convolution theorem for Gaussians [7]. It is fundamental to note that there are no approximations in the calculation of

$\int \hat{p}(x)\hat{q}(x)dx$. Furthermore the final expression (4) depends exclusively on observations x_i.

In a similar manner one may get

$$\int p^2(x)dx = \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} G(x_i - x_j, 2\sigma^2), \qquad (5)$$

and also

$$\int q^2(x)dx = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(x_i - x_j, 2\sigma^2). \qquad (6)$$

By applying (4), (5) and (6) in (3), one gets

$$D_{C\text{-}S}(p,q) = \log\int p^2(x)dx - 2\log\int p(x)q(x)dx + \log\int q^2(x)dx$$

$$\approx \log\frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} G(x_i - x_j, 2\sigma^2)$$

$$- 2\log\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} G(x_i - x_j, 2\sigma^2)$$

$$+ \log\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(x_i - x_j, 2\sigma^2).$$

Acknowledgements

Carlos Pedreira was partially supported by grants from CNPq (Brazilian National Research Council) and FAPERJ (Rio de Janeiro Research Foundation). Rodrigo Peres was supported by a Ph.D. scholarship from CNPq.