New Results on Observations Selection

R. T. PERES[1] AND C. E. PEDREIRA[2]
[1]Electrical Engineering Department – PUC-Rio.
[2]Faculty of Medicine and the Engineering Graduate Program (COPPE - PEE), Federal University of Rio de Janeiro, UFRJ.
BRAZIL

Abstract: - In this paper we propose a new observations selection technique in Vector Quantization context. The main idea is to select observations that are not representatives of their classes. The cells generated by the Vector Quantization are divided in regions of rejection. A number of controlled experiments were performed demonstrating the proposed methodology potentiality.

Key-Words: - Observations Selection, Vector Quantization, Classification.

1 Introduction

In this paper we are concerned with observations selection. To use the entire available set of observations may not always be the best strategy. In many situations, it is interesting to select a subset of the sample. This selection may produce a set of observations that are more representative of their classes and consequently improve the performance in classification problems. In a quite different perspective, observations selection could be also associated to outliers identification.

There are a number of previous contributions in the literature concerning observations selection under a variety of approaches: Active Learning [14], [5] demands a previous specification of the model and parameters. So, the associated observations selection is dependent on the quality of this specification [9]. The method proposed in [7] groups the observations in three categories: typical, critical and noisy by considering the intrinsic margin, a measure of the distance between observation and the decision boundary. A method to identify outliers, based on exhaustive learning can be found in [10]. [9] focused on observations selection from a Bayesian perspective. Other strategies are query-based learning [6] and sequential design [2]. [12] propose the risk-zone concept in a Learning Vector Quantization (LVQ) context. The key idea is to select a subset of observations with the goal of conducting the prototypes to convenient locations other than the class mean. This methodology was successfully applied in a heart diseases diagnosis problem [11]. Also, in a Vector Quantization (VQ) context, [13] proposed a method where a discriminating mapping is applied to select observations after projecting them in a bi-dimensional space.

Here, we propose an observations selection methodology in a supervised environment. The main goal is to identify, and possibly exclude, observations that are considered to be unrepresentative of their classes.

2 Methodology

Let us consider a dichotomous classification environment where $X = \{x_1, x_2, ..., x_n\}$ is a set of observations (each $x_i \in R^p$, $\forall i = 1,...,n$). We assume that each observation x_i belongs to one of two classes C_1 and C_2, with associated labels 0 or 1 respectively. The methodology that follows can be extended to multiple classes problems in the usual manner [1]. The objective is two fold: to identify observations that are not representatives of their classes or that have had their labels inverted by some noise mechanism.

Let us denote $y(x_i)$ as the label of observation x_i, i.e., $y(x_i) = 0$ if x_i belongs to C_1, and $y(x_i) = 1$ if x_i belongs to C_2. It is assumed that the assigned labels, in the dataset, may eventually not correspond to the true label due to action of a noise mechanism. Because of that we define $\Im(x_i)$ as the true label associated to the observation x_i. The concept of inversion of label can be now established as follows: An observation x_i has had its label noisily inverted if $y(x_i) \neq \Im(x_i)$.

Vector Quantization (VQ) [4], [3], that has been extensively explored in literature, is used here as a first step. The main idea is to establish a quantized approximation of the data distribution, using a finite number of prototypes. These prototypes may be associated with the observations by the nearest neighbor rule.

Let $P = \{p_1, ... , p_r\}$, $p_k \in R^p$, $\forall k = 1, ..., r$ be a set of prototypes. A VQ procedure can be defined as an association of each observation $x_i \in X$ to a prototype p_k. In general, this association is done by linking the observations to their nearest prototype according to some specific metric. In this paper we use the Euclidian distance, but other metrics may be more convenient to some specific applications and methods e.g. [12]. As a result of the quantization procedure, the set of observations X ends up partitioned in subsets that we call cells. These cells are denoted by $S_1,..., S_r$. Formally:

$S_k \equiv \{ x_i \in X \mid d(x_i, p_k) \leq d(x_i, p_j), j=1,...,r; j\neq k \}, \forall k = 1,..., r.$

Here, we use the LBG algorithm [8] for quantization step, with the goal of segmenting the sample in cells. These cells will be individually treated in next sections. The LBG is an unsupervised procedure, and so, it is possible that some cells end up constituted by a heterogeneous population concerning the labels of the observations. The LBG algorithm is initiated with 2 prototypes, generating two cells. Next, these prototypes are repeatedly updated to the center of the cells until the average distortion variation lays below a threshold ς. After that, each prototype p_k is substituted by $p_k + \lambda$ and $p_k - \lambda$ (where λ is a small valued parameter). This procedure is repeated until a maximum pre-established number of prototypes are reached.

For each cell S_k, generated by the VQ step, we define the following two sets: $W_k \equiv \{x \in S_k \mid y(x) = 0\}$ and $T_k \equiv \{x \in S_k \mid y(x) = 1\}$, and determine the correspondent frequencies of classes C_1 and C_2:

$$f_0(k) = \frac{\#W_k}{\#S_k} \quad \text{and} \quad f_1(k) = \frac{\#T_k}{\#S_k}$$

where #A represents the cardinality of a set A. Since we are interested in non-representative observations, we restrain our attention to the heterogeneous cells. Note that $W_k = \varnothing \Rightarrow f_0(k) = 0$ and $T_k = \varnothing \Rightarrow f_1(k) = 0$. In these cases the cell is fully homogeneous and consequently skipped.

Given a cell S_k, we denote its highest frequency as $f\_max \equiv \arg\max_{0,1}(f_0(k), f_1(k))$.

Let us consider the means of each class in a cell S_k:

$$m_0(k) = \frac{1}{\#W_k} \sum_{x \in W_k} x \quad \text{and} \quad m_1(k) = \frac{1}{\#T_k} \sum_{x \in T_k} x$$

Definition 1: We say that an observation $x \in S_k$ belongs to the rejection region of class C_1, namely $R_0(k)$, if

$$\frac{d(x, m_0(k))}{d(x, m_1(k))} > \Omega_1 .$$

Analogously, an observation $x \in S_k$ belongs to the rejection region of class C_2, $R_1(k)$, if

$$\frac{d(x, m_1(k))}{d(x, m_0(k))} > \Omega_2.$$

The thresholds are defined as $\Omega_1 \equiv \frac{f_0(k)}{f_1(k)}$ and $\Omega_2 \equiv \frac{f_1(k)}{f_0(k)}$. In other words, the rejection regions of classes C_1 and C_2 are defined as

$R_0(k) \equiv \{x \in S_k \mid \frac{d(x, m_0(k))}{d(x, m_1(k))} > \Omega_1 \}$ and $R_1(k) \equiv \{x \in S_k \mid \frac{d(x, m_1(k))}{d(x, m_0(k))} > \Omega_2\}$.

Note that, unless in the improbable case of equality, $R_0(k)$ and $R_1(k)$ are complementary regions, and so, they represent a partition of S_k.

Definition 2: An observation $x \in S_k$ is selected if: (i) $x \in R_0(k) \wedge y(x) = 0$; (ii) $x \in R_1(k) \wedge y(x) = 1$.

If W_k or T_k are unitary, the unique observation is automatically selected, i.e., if $\#W_k = 1$, $R_0(k)=\{x_1\}$ or if $\#T_k = 1$, $R_1(k)=\{x_1\}$, the observation x_1 is selected.

Note that if $\Omega_1 = \Omega_2 = 1$, the procedure selects the observations that are near the other class mean, since $R_0(k) \equiv \{x \in S_k \mid d(x, m_0(k)) > d(x, m_1(k))\}$ and $R_1(k) \equiv \{x \in S_k \mid d(x, m_1(k)) > d(x, m_0(k))\}$. The process described is repeated for all cells S_k, k = 1,..,r.

The key idea is to select observations for which the rate of the distances between the observation to the mean of its class and to the mean of the other class to exceed a, frequency-based, threshold value (Ω_1 or Ω_2, depending of the label of the observation). In this way, the sizes of the rejection regions vary in accordance to the measured frequency of each class in a given cell.

The observations that belong to the rejection regions $R_0(k)$ or $R_1(k)$ are selected as unrepresentative of their classes. These observations may have been generated by some sort of noise mechanism resulting in label inversion. Taken the rate of frequencies in a given cell, the selected observations are relatively far away from the mean of their class (in relation to their distance to the mean of the other class).

Note that, since we are dealing with a dichotomous classification environment, the frequencies in each class are complementary. So, $f_0(k) + f_1(k) = 1$, $\forall k = 1,…,r$. Also, in a situation of complete equilibrium, $f_0(k) = f_1(k) = 0.5$.

We detach two situations of approximately equilibrium: observations are disorderly mixed, as in

Fig. 1, or in two distinct groups, as in Fig. 2 (possibly characterized by a cell in the boundary decision). First case is hopeless and we opt to select (and possibly discard) all the observations of this cell. We distinguish these cases as follows: if f_max $\leq \alpha$ and $d(m_0(k), m_1(k)) < \beta$ for chosen thresholds α and β, the observations are disorderly mixed.

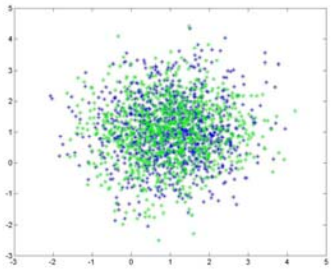

Fig. 1. A cell where the classes have the same frequency and the observations are mixed.

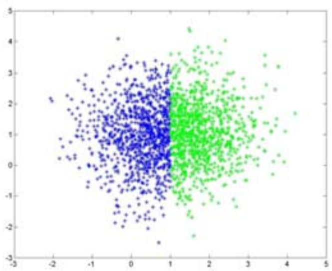

Fig. 2. A cell where the classes have the same frequency but observations are divided in two groups.

3 Results

The following parameters were used for all experiments: thresholds $\alpha = 0.6$ and $\beta = 0.5$; $\lambda = 10^{-1}$ and $\varsigma = 10^{-2}$ (LBG algorithm).

Experiment 1: We built up a dataset consisting of two classes divided by a cosine function (2070 and 2053 observations for classes C_1 and C_2 respectively). For each class, 20 observations had their label deliberately inverted (nearly 1% of the observations), i.e. for these observations, $y(x_i) \neq \Im(x_i)$. These inverted label observations were uniformly distributed. In Fig. 3, we represent C_1 and C_2 observations with asterisks. The inverted label observations are marked as circles. Results are shown in table 1. By erroneously selected observations, we mean observations that are not inverted labels ones but were mistakenly selected as so.

Table 1. Performance (Experiment 1).

Classes	Identified Inverted Labels	Erroneously Selected Observation
C_1	19 out of 20 (95%)	37
C_2	18 out of 20 (90%)	41
Total	37 out of 40 (92.5%)	78 out of 4083 (1.9%)

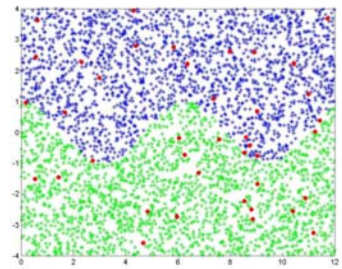

Fig. 3. Experiment 1 - C_1 and C_2 as asterisks and observations with label inverted as circles.

Experiment 2: The dataset is generated by a circle and a roll with the same centers and no intersection (123 observations belonging to C_1 and 2611 observations to C_2), see Fig. 4. The labels of 5 observations inverted in C_1 and of 20 observations in C_2. Results are presented in table 2.

Table 2. Performance (Experiment 2).

Classes	Identified Inverted Labels	Erroneously Selected Observation
C_1	4 out of 5 (80%)	28
C_2	20 out of 20 (100%)	11
Total	24 out of 25 (96%)	39 out of 2734 (1.4%)

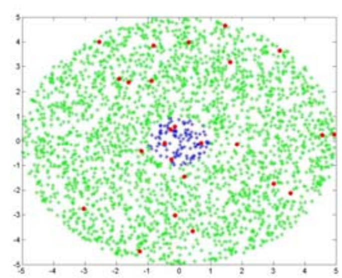

Fig. 4. Experiment 2 - C_1 and C_2 as asterisks and observations with label inverted as circles.

Experiment 3: This dataset was built up with two classes (1000 observations belonging to C_1 and 1000 observations to C_2), uniformly distributed, as squares with 25% of the area in common and we introduced 20 observations with inverted labels for each class (Fig. 5a). Results are in table 3. The method selected 475 observations (246 from C_1 and 229 from C_2) with the majority of the central square plus the 40 inverted label ones. These observations were removed in Fig. 5b.

Table 3. Performance - Experiment 3

	Inverted Labels Detected
C_1	20 out of 20 (100%)
C_2	20 out of 20 (100%)
Total	40 out of 40 (100%)

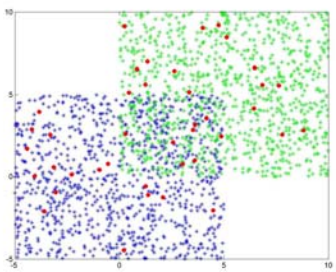

Fig. 5a. Experiment 3 - C_1 and C_2 as asterisks, observations with label inverted as circles and the uniform area in the center.

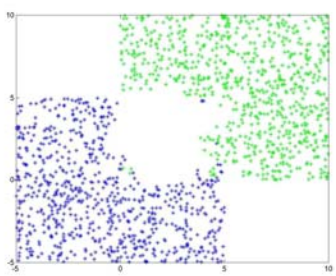

Fig. 5b. Experiment 3 without observations selected by the method.

4 Conclusion

In this paper we described a new methodology for data selection. After an unsupervised step, (VQ), the procedure is used to identify observations that are not representative of their classes. Each heterogeneous cell obtained by the quantization is divided in regions of rejection for each class. Observations that belong to the rejection region of its class are selected by the method. The potentiality of the proposed methodology was positively evaluated through three synthetically experiments. The methodology was clearly efficient in terms of recognizing observations with inversion of label.

References:

[1] Duda, R.O., Hart, P.E., Stork, G, *Pattern Recognition*, 2nd. Ed., Wiley, US, 2001.

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

[3] Gersho, A., Gray, R.M., *Vector Quantization and Signal Compression*, Kluwer Academic Press/Springer, US, 1992.

[4] Gray, R. M., Vector quantization, *IEEE ASSP Magazine*, Vol. 1, Issue 2, 1984, pp. 4-29.

[5] Hasenjäger, M., Ritter, H., Obermayer, K., Active learning in self-organizing maps, *In E.Oja & S. Kaski editors*, Kohonen Maps, Elsevier, 1999, pp. 57-70.

[6] Hwang, J. N., Choi, J. J., Oh, S., Marks II, R. J., Query-based learning applied to partially trained multi-layer perceptrons, *IEEE Trans. Neural Networks*, Vol.2, Issue 1, 1991, pp. 131-136.

[7] Li, L., Pratap, A., Lin, H.-T., Abu-Mostafa, Y. S., Improving generalization by data categorization, *PKDD, LNAI 3721*, Springer-Verlag, 2005, pp. 157-168.

[8] Linde, Y., Buzo, A., Gray, R. M., An algorithm for vector quantizer design, *IEEE Trans. Communications*, Vol. COM-28, Issue 1, 1980, pp. 84-95.

[9] MacKay D. J. C., Information-based objective functions for active data selection, *Neural Computation*, Vol. 4, Issue 4, 1992, pp. 590-604.

[10] Nicholson, A., *Generalization error estimates and training data valuation*, Ph.D. Dissertation, California Institute of Technology, US, 2002.

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

[12] Pedreira, C. E., Learning vector quantization with training data selection, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 28, Issue 1, 2006, pp. 157-162.

[13] Peres, R. T., Pedreira, C.E., Preliminary Results on Noise Detection and Data Selection for Vector Quantization, *Proceedings of IEEE World Congress on Computational Intelligence*, Vancouver, 2006.

[14] Plutowski, M., White, H., Selecting concise training sets from clean data, *IEEE Trans. Neural Networks*, Vol. 4, Issue 2, 1993, pp. 305-318.

Acknowledgment: This work has been partially supported by grants from the **CNPq-** Brazilian National Research Council and **FAPERJ** – Rio de Janeiro Research Foundation, Brazil.