Detecting uncertainty regions for characterizing classification problems

Gian Paolo Drago, Marco Muselli

Istituto per i Circuiti Elettronici - CNR
via De Marini, 6 - 16149 Genova, Italy

Abstract

A mathematical framework for the analysis of critical zones of the input space in a classification problem is introduced. It is based on the definition of uncertainty region, which is the collection of the input patterns whose classification is not certain. Through this definition a characterization of optimal decision functions can be derived.

A general method for detecting the uncertainty region in real-world problems is then proposed, whose implementation can vary according to the connectionist model employed. Its application allows to improve the performance of the resulting neural network.

1 Introduction

Generalization ability is the quantity usually employed to measure the quality of a connectionist model, which has been trained to solve a real-world classification problem. The value of this quantity is normally obtained by observing the performance of the neural network on a validation set, containing some samples pertaining to the given classification problem and not used in the training process.

However, generalization ability gives only an overall evaluation of the effectiveness of our artificial device and does not shed much light on its local properties or on the characteristics of the real-world problem at hand. In particular, no information is provided about the quality of the classification for a specific input pattern; consequently, we cannot know if the point lies in a critical region of the input space or if its class can be assigned with a high confidence.

This knowledge is not only important when the trained connectionist model is used; the learning process can indeed search more carefully critical zones of the input space so as to achieve a better final performance. Unfortunately, no way of formally characterizing these critical zones is available in the literature; only some heuristic methods is given [4, 5, 2], whose validity is not generally ensured.

In the following sections a mathematical framework for two-class problems is introduced, which allows to define the concept of *uncertainty region* U as the collection of the input patterns whose classification is not certain. A characterization of optimal decision functions based on the uncertainty region U is then introduced. However, the practical application of this framework requires

the ability of computing U in a real-world pattern recognition problem. To this aim a general algorithm is presented, whose implementation may differ according to the connectionist model considered. A version based on Support Vector Machines is presented in a companion paper [3].

2 The mathematical framework

Consider a general pattern recognition problem, where vectors $x \in \mathcal{R}^d$ have to be assigned to one of two possible classes, associated with the values of a binary output y, coded by the integers -1 and $+1$. Following the mathematical model of [1], this problem can be thoroughly described by a pair of probability measures (μ, η), where μ is defined on the Borel sets of $\mathcal{R}^d$ and η is the so-called *a posteriori probability* given by $\eta(x) = \mathbf{P}\{Y = +1 \mid X = x\}$. X and Y are random variables assuming values in $\mathcal{R}^d$ and $\{-1, +1\}$, respectively.

It can be easily seen that the conditional probability $\eta(x)$ contains all the necessary information about the pattern recognition problem at hand. The goal is to find a *decision function* (also called *classifier*) $g : \mathcal{R}^d \to \{-1, +1\}$ that minimizes the error probability $L(g) = \mathbf{P}\{g(X) \neq Y\}$. A one-one correspondence between the class of decision functions and the collection of Borel subsets of $\mathcal{R}^d$ is directly obtained; for example, we can associate with any classifier g the set $D^+(g)$ containing all the points x of the input space $\mathcal{R}^d$ for which $g(x) = +1$. A corresponding subset $D^-(g)$ including the portion of $\mathcal{R}^d$ where $g(x) = -1$ can also be introduced.

$$D^+(g) = \{x \in \mathcal{R}^d \,:\, g(x) = +1\}\,, \qquad D^-(g) = \{x \in \mathcal{R}^d \,:\, g(x) = -1\}$$

The boundary between $D^-(g)$ and $D^+(g)$, defined as $B(g) = \mathrm{cl}\, D^-(g) \cap \mathrm{cl}\, D^+(g)$, being $\mathrm{cl}\, A$ the closure of A, will be called *separating set* of the classifier g.

Two decision functions g and g' can be considered equivalent if $g = g'$ a.s., that is $\mu\{g(X) \neq g'(X)\} = 0$; in this case we write $g \sim g'$. Thus, equivalence classes in the set of classifiers can be determined, according to the given measure probability μ acting on the input space $\mathcal{R}^d$. It is straightforward to see that $L(g) = L(g')$ when $g \sim g'$.

Let $\mathcal{G}$ be the set containing one measurable decision function g for each of the equivalence classes just introduced. If $\eta(x)$ were known, the optimal classifier for our pattern recognition problem is the *Bayes decision function* $g^* \in \mathcal{G}$, defined as $g^*(x) = +1$, if $\eta(x) > 1/2$, and $g^*(x) = -1$ otherwise. The error probability $L^* = L(g^*)$ has been shown to be minimal [1], that is $L^* \leq L(g)$ for any decision function $g \in \mathcal{G}$.

Two subsets C^- and C^+ in the input space $\mathcal{R}^d$ may be determined, according to the value assumed by the a posteriori probability $\eta(x)$

$$C^- = \{x \in \mathcal{R}^d \,:\, \eta(x) = 0\}\,, \qquad C^+ = \{x \in \mathcal{R}^d \,:\, \eta(x) = 1\}$$

they will be called *certainty regions*, since the points belonging to C^- and C^+ are univocally associated with a specific output. For a similar reason, the

remaining portion of the input space $U = \mathcal{R}^d \setminus (C^- \cup C^+)$ will be named *uncertainty region*.

The knowledge of the subsets C^-, C^+, and U gives important insights into the classification problem at hand. In fact, it will be shown that every valid classifier g must assume value -1 in C^- and value $+1$ in C^+. To see this, let us introduce the class of functions $\bar{\mathcal{G}} \subset \mathcal{G}$ containing this kind of classifiers

$$\bar{\mathcal{G}} = \{g \in \mathcal{G} \,:\, C^+ \subset D^+(g), C^- \subset D^-(g)\}$$

Then, a proper transformation $T : \mathcal{G} \to \bar{\mathcal{G}}$ can be defined

$$T(g)(x) = \begin{cases} -1 & \text{if } x \in C^- \\ +1 & \text{if } x \in C^+ \\ g(x) & \text{if } x \in U \end{cases}$$

Note that $\bar{\mathcal{G}}$ is invariant under this transformation, since $T(g) = g$ for every $g \in \bar{\mathcal{G}}$. With these definitions it will be shown that best classifiers are included in $\bar{\mathcal{G}}$.

Theorem 1 *For every decision function $g \in \mathcal{G} \setminus \bar{\mathcal{G}}$ we have $L(T(g)) < L(g)$.*

Proof. The conditional error probability of any classifier $g \in \mathcal{G}$, given $X = x$, can be written as [1]

$$\begin{aligned} &\mathbf{P}\{g(X) \neq Y \mid X = x\} = 1 - \mathbf{P}\{g(X) = Y \mid X = x\} \\ &= 1 - \left(I_{D^+(g)}\mathbf{P}\{Y = +1 \mid X = x\} + I_{D^-(g)}\mathbf{P}\{Y = -1 \mid X = x\}\right) \\ &= 1 - \left(I_{D^+(g)}\eta(x) + I_{D^-(g)}(1 - \eta(x))\right) \end{aligned}$$

being I_A the indicator function of the set A. Thus, for every $x \in \mathcal{R}^d$, we have

$$\begin{aligned} &\mathbf{P}\{g(X) \neq Y \mid X = x\} - \mathbf{P}\{T(g)(X) \neq Y \mid X = x\} \\ &= \eta(x)\left(I_{D^+(T(g))} - I_{D^+(g)}\right) + (1 - \eta(x))\left(I_{D^-(T(g))} - I_{D^-(g)}\right) \\ &= I_{D^-(g)}I_{C^+} + I_{D^+(g)}I_{C^-} \end{aligned}$$

since $T(g)(x) = g(x)$ when $x \in U$. By definition of $\mathcal{G}$, this allows to conclude that $L(g) - L(T(g)) = \mu(C^+ \cap D^-(g)) + \mu(C^- \cap D^+(g)) > 0$ if $g \in \mathcal{G} \setminus \bar{\mathcal{G}}$. □

As a consequence of Theorem 1 we obtain that the Bayes classifier g^* lies in the class $\bar{\mathcal{G}}$; otherwise, the decision function $T(g^*)$ would achieve a lower error probability. Furthermore, it is possible to determine two classifiers $g^-, g^+ \in \bar{\mathcal{G}}$ such that $D^-(g^-) = C^-$ and $D^+(g^+) = C^+$ a.s. Due to the above defined one-one correspondence between the class of decision functions and the collection of subsets of $\mathcal{R}^d$, the introduction of g^- and g^+ can be viewed as an alternative definition of the sets C^- and C^+, which leads to $U = D^+(g^-) \cap D^-(g^+)$.

3 A general algorithm for detecting U

When solving real world pattern recognition problems, usually we do not know the probability measures μ and η, but have only access to a training set S_n containing n samples (X_j, Y_j), $j = 1, \ldots, n$, supposed to be obtained through n i.i.d. applications of μ and η. Consequently, the problem of determining the regions C^-, C^+, and U cannot be solved unless some kind of regularity is imposed on the choice of the decision functions g^- and g^+ [6].

To this aim, let us introduce a hypothesis set $\mathcal{H}$, containing all the classifiers g among which the desired behaviors for g^- and g^+ are searched for. Since the decision functions g^- and g^+ may not belong to the class $\mathcal{H}$, our target is therefore to determine two classifiers $\hat{g}^-, \hat{g}^+ \in \mathcal{H}$, such that $D^-(\hat{g}^-) \subset C^-$ and $D^+(\hat{g}^+) \subset C^+$, while the portion of C^- and C^+ contained in $D^+(\hat{g}^-)$ and $D^-(\hat{g}^+)$, respectively, is minimal, according to the unknown probability measure μ. Formally, we have

$$\begin{aligned} \hat{g}^- &= \arg\min_{g \in \mathcal{H}} \{\mu(D^+(g) \cap C^-) \,:\, D^-(g) \subset C^-\} \\ \hat{g}^+ &= \arg\min_{g \in \mathcal{H}} \{\mu(D^-(g) \cap C^+) \,:\, D^+(g) \subset C^+\} \end{aligned}$$

A possible way to determine through the training set S_n the behavior of the decision functions $\hat{g}^-$ and $\hat{g}^+$ is to perform the following couple of maximizations

$$\max_{g \in \mathcal{H}_n^-} \left|S_n^- \cap D^-(g)\right| \,, \qquad \max_{g \in \mathcal{H}_n^+} \left|S_n^+ \cap D^+(g)\right| \tag{1}$$

where $|A|$ is the *cardinality* of the set A (i.e. the number of elements in A when it is finite), whereas the sets $\mathcal{H}_n^-$, $\mathcal{H}_n^+$, S_n^-, and S_n^+ are defined as follows:

$$\begin{aligned} \mathcal{H}_n^- &= \{g \in \mathcal{H} \,:\, D^-(g) \cap S_n^+ = \emptyset\} \\ \mathcal{H}_n^+ &= \{g \in \mathcal{H} \,:\, D^+(g) \cap S_n^- = \emptyset\} \\ S_n^- &= \{X \in \mathcal{R}^d \,:\, (X, Y) \in S_n,\, Y = -1\} \\ S_n^+ &= \{X \in \mathcal{R}^d \,:\, (X, Y) \in S_n,\, Y = +1\} \end{aligned}$$

The goal of the first term in (1) is to determine the decision functions $g \in \mathcal{H}$ that maximize the number of correctly classified patterns X in the training set S_n, having corresponding output $Y = -1$, while satisfying all the samples $(X, Y) \in S_n$ with $Y = +1$. For notational purposes we have therefore introduced the class $\mathcal{H}_n^-$ containing all the decision functions $g \in \mathcal{H}$ that give output $g(X) = +1$ in correspondence of the elements of the set S_n^+; this is formed by the input patterns X belonging to the training set S_n with an associated $Y = +1$.

Similar definitions for $\mathcal{H}_n^+$ and S_n^- have also been employed in the determination of the classifier $\hat{g}^+$. Actually, maximizations (1) may not lead to a unique solution, but several decision functions of $\mathcal{H}$ may achieve the same maximum number of correctly classified patterns. Let $\bar{\mathcal{H}}_n^-$ and $\bar{\mathcal{H}}_n^+$ be the two subsets of $\mathcal{H}$ containing the possible arguments of maxima in (1).

A further choice among the elements of $\bar{\mathcal{H}}_n^-$ and $\bar{\mathcal{H}}_n^+$ is therefore necessary to obtain good approximations $\hat{g}_n^-$, $\hat{g}_n^+$ for the desired decision functions $\hat{g}^-$ and $\hat{g}^+$. This choice can be made in the following way, although several valid alternatives exist:

$$\hat{g}_n^- = \arg \max_{g \in \bar{\mathcal{H}}_n^-} \lambda(D^-(g)) \,, \qquad \hat{g}_n^+ = \arg \max_{g \in \bar{\mathcal{H}}_n^+} \lambda(D^+(g))$$

being $\lambda(A)$ the Lebesgue measure of the set $A \subset \mathcal{R}^d$. In this way the classifiers $\hat{g}_n^-$ and $\hat{g}_n^+$ are selected, which maximize the extension (in the sense of Lebesgue measure) of the sets $D^-(g)$ and $D^+(g)$, respectively.

Now, the approximations C_n^-, C_n^+, and U_n for the desired sets C^-, C^+, and U can be obtained through the equations

$$\begin{aligned} C_n^- &= D^-(\hat{g}_n^-) \cap D^-(\hat{g}_n^+) \,, \qquad C_n^+ = D^+(\hat{g}_n^+) \cap D^+(\hat{g}_n^-) \\ U_n &= \mathcal{R}^d \setminus (C_n^- \cup C_n^+) \end{aligned}$$

Summing up, the procedure to be employed for obtaining from the given training set S_n the estimates C_n^-, C_n^+, and U_n can be outlined as in Fig. 1.

PROCEDURE FOR APPROXIMATING C^-, C^+, AND U

1. Determine the sets of decision functions $\bar{\mathcal{H}}_n^-$ and $\bar{\mathcal{H}}_n^+$, whose elements achieve the maxima in (1).
2. Choose in $\bar{\mathcal{H}}_n^-$ and in $\bar{\mathcal{H}}_n^+$ the classifiers $\hat{g}_n^-$ and $\hat{g}_n^+$, which maximize the Lebesgue measure of $D^-(g)$ and $D^+(g)$, respectively.
3. Set $C_n^- = D^-(\hat{g}_n^-) \cap D^-(\hat{g}_n^+)$, $C_n^+ = D^+(\hat{g}_n^+) \cap D^+(\hat{g}_n^-)$, and $U_n = \mathcal{R}^d \setminus (C_n^- \cup C_n^+)$.

Figure 1: General procedure followed to obtain reasonable approximations for the sets C^-, C^+, and U from a finite training set S_n.

4 Discussions

It is interesting to note that if the training set S_n is perfectly separable by a decision function $g \in \mathcal{H}$, i.e. $S_n^- \subset D^-(g)$ and $S_n^+ \subset D^+(g)$, the estimated uncertainty region U_n includes the largest region of the input space, realizable through the classifiers in $\mathcal{H}$, which does not contain patterns of S_n. In this case, the separating set $B(g)$ of every decision function $g \in \mathcal{H}$ that correctly classifies all the patterns in the given training set S_n is contained in $\mathrm{cl}\, U_n$.

However, in a general case Theorem 1 ensures that the separating set of any good classifier lies into $\mathrm{cl}\, U$. Thus, if we are confident with our approximations C_n^- and C_n^+, obtained through S_n by employing the procedure in Fig. 1, the

solution of the classification problem at hand can be pursued by examining only the behavior inside the uncertainty region U_n. Consequently, the samples of S_n included in C_n^- and C_n^+ may be removed from the training set, without compromising the search for the target decision function. This leads to a reduction of the computational burden needed for the application of any pattern recognition algorithm.

Furthermore, if new samples are to be added to the actual training set, we should discard all the input patterns belonging to C_n^- and C_n^+, since their presence does not bring essential information for the problem at hand. In particular, if an oracle is available, which gives the corresponding output y for any selected pattern x, we can consider for a possible inclusion in the training set only the points of the uncertainty region U_n.

If rejections are admissible in the application we are dealing with, one can adopt the following decision function with reject option

$$g(x) = \begin{cases} -1 & \text{if } x \in C_n^- \\ +1 & \text{if } x \in C_n^+ \\ \text{“reject”} & \text{if } x \in U_n \end{cases}$$

as the solution of the classification problem at hand. In this case, no further computation is required to achieve a final two-valued decision function. Naturally, the size of the uncertainty region U_n must be small enough to keep low the rejection rate.

References

[1] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition.* New York: Springer-Verlag (1997).

[2] G. P. Drago and S. Ridella, Possibility and necessity pattern classification using an interval arithmetic perceptron. *Neural Computing & Applications*, **8** (1999), 40–52.

[3] G. P. Drago and M. Muselli, Support Vector Machines for uncertainty region detection. Submitted to the *12-th Italian Workshop on Neural Nets* (2001).

[4] H. Ishibuchi, R. Fujioka, and H. Tanaka, Possibility and necessity pattern classification using neural networks. *Fuzzy Sets and Systems*, **48** (1992), 331–340.

[5] H. Ishibuchi, H. Tanaka, and H. Okada, An architecture of neural networks with interval weights and its application to fuzzy regression analysis. *Fuzzy Sets and Systems*, **57** (1993), 27–39.

[6] V. N. Vapnik, *Statistical Learning Theory.* New York: John Wiley & Sons (1998).