"ACCELERATED PERCEPTRON": A SELF-LEARNING LINEAR DECISION ALGORITHM

Zuev Yu.A.

Computing Center of Russion Academy of Sciences, Vavilova 40 119991 Moscow, Russia

Abstract. The class of linear decision rules is studied. A new algorithm for weight correction, called an "accelerated perceptron", is proposed. In contrast to classical Rosenblatt's perceptron this algorithm modifies the weight vector at each step. The algorithm may be employed both in learning and in self-learning modes. The theoretical aspects of the behaviour of the algorithm are studied when the algorithm is used for the purpose of increasing the decision reliability by means of weighted voting. In this case the simple majority vote may be used as initial decision.

1. Introduction

Linear decision rules are widely spread in pattern recognition when features are quantified and there are 2 classes: $\mathbf{K}_1$ and $\mathbf{K}_2$. In this case an object $\mathbf{Q}$ is described by n-dimension vector $(x_1,\mathrm{K}\,,x_n)$. The linear decision rule is determined by n-dimension weight vector $(w_1,\mathrm{K}\,,w_n)$ and threshold w_0. The object is classified into class $\mathbf{K}_1$ if $w_1x_1+\mathrm{K}+w_nx_n>w_0$ and into class $\mathbf{K}_2$ if $w_1x_1+\mathrm{K}+w_nx_n<w_0$. It is useful to introduce $(n+1)$-dimension vector $X=(-1,\ x_1,\mathrm{K}\,,x_n)$ and $(n+1)$-dimension vector $W=(w_0,w_1,\mathrm{K}\,,w_n)$. Then the object is classified into class $\mathbf{K}_1$, if $\mathrm{sgn}(W,X)=1$ and into class $\mathbf{K}_2$ if $\mathrm{sgn}(W,X)=1$. The problem is to find a vector W which provides a good discrimination among two classes.

Let first consider the learning case when we have a training set consisting of some (for a example l) objects from the class $\mathbf{K}_1$ and some (for a example m) objects from the class $\mathbf{K}_2$. In such case we may try to find W which provides minimal number of erroneous discriminated objects. If l and m are large enough ($l,\ m\ >>\ n$) this approach by all means will provide good discrimination but it may take anomalous amount of computing work. On the other hand if l and m are rather small it may be not so difficult to find W which separates training objects without any mistakes but there is a good deal of uncertainty in choosing such W. In this case we would like W, which maximizes distances between training objects and separating hyperplane.

Point out a simple recipe to overcome these difficulties. Let us determine the central points of training sets in classes $\mathbf{K}_1$ and $\mathbf{K}_2$ as $X_1^\beta=\frac{1}{l}\sum_{X_i\in K_1}X_i$, $X_2^\beta=\frac{1}{m}\sum_{X_j\in K_2}X_j$ and define the vector $W^\beta=X_1^\beta-X_2^\beta$. Further this weight vector and determined by it decision rule will be called *baricentric*. In many cases baricentric rule may be practically acceptable. But for many years an attention of investigators working in the field of pattern recognition was concentrated on another approach consisting in sequential recurrent weight modifying. This approach known as *perceptron* was firstly suggested by Rosenblatt [5].

2. Drawbacks of classical perceptron

Let's connect with each object $\mathbf{Q}_i$ from the training set a variable z_i: $z_i=1$, if $\mathbf{Q}_i\in\mathbf{K}_1$ and $z_i=-1$, if $\mathbf{Q}_i\in\mathbf{K}_2$. The objects $\mathbf{Q}_i$ are picked out of the training set

sequentially and the weight vector is modified at each step by classical perceptron algorithm as follows:

$$W_{k+1} = \begin{cases} W_k, & if \quad \mathrm{sgn}(W_k, X_k) = z_k \\ W_k + z_k X_k, & if \quad \mathrm{sgn}(W_k, X_k) \neq z_k \end{cases}.$$

The classical perceptron algorithm was studied in a number of works [1,2,3,5].A remarkable property of it is the following. If the objects are picked up cyclically out of the training set then after a finite number of weights corrections the algorithm finds the hyperplane correctly separating the training set, if such a hyperplane exists. In practice, however, nothing at all is usually known about separability. At the same time, classical perceptron a number of drawbacks. These include:

1. slow learning, since the weight vector is only corrected when misclassification takes place;
2. random final position of the separating hyperplane in the case of linear separability, of the training set while common sense and experience suggest that good separation requires the training objects to be as far from the hyperplane as possible;
3. in the case of linear inseparability of the training set, the modulus of the weight vector remains small during learning and there is a large change in hyperplane position at each correction, leading to strong fluctuations in the quality of decision rule;
4. self-learning is impossible.

These drawbacks led to striking the perceptron off the list of practically working algorithms. Our aim is to show that after a little modification the perceptron idea may be done practically working.

3. "Accelerated perceptron"

In self-learning case we have a set of objects from two classes K_1 and K_2 too, but it is unknown which object from which class is, that is z_i are unknown. If we have some initial decision rule satisfactory quality, we can increase it in this case too. The idea formulated by Pierce [4] implies using the decision rule output instead of unknown z_i. In other words the output of decision rule is considered as the true value of z_i. Classical perceptron is unable to self-learning because it doesn't modified the weight vector when true classification taking place.

The algorithm "accelerated perceptron" [6, 7] modifies the weight vector at each step, regardless true or erroneous classification taking place. That's why it can be employed in self-learning mode using Pierce's principle. The accelerated percepron does not suffer from the drawbacks of classical perceptron, although it does not guarantee error-free separation in the separate case.

The weight correction rule for an accelerated perceptron in learning mode is

$$W_{k+1} = W_k + z_k X_k.$$

The weight correction rule for an accelerated perceptron in self-learning mode is

$$W_{k+1} = W_k + \mathrm{sgn}(W_k, X_k) X_k.$$

It's not difficult to see that in learning mode the weight vector tends to become collinear to baricentric vector and decision rule tends to become baricentric. In self-learning mode the behavior of the accelerated perceptron is more complicated. For theoretical

analysis of its behavior in this case we'll consider a special problem of optimal voting decision.

4. The problem of optimal committee decision

Let's we have as before a pattern recognition problem with two classes $\mathbf{K}_1$ and $\mathbf{K}_2$ and n classifiers which classifier presented object $\mathbf{Q}$ in one of two classes. Classifiers may be experts, technical systems or pattern recognition algorithms. The decision i-th classifier $y_i = 1$, if it classifiers an object into class $\mathbf{K}_1$ and $y_i = -1$ in opposite case. The true belonging of the object is $z \in \{-1,\ 1\}$. The problem is to classify the object on the basis of n-dimension vector $Y = (y_1, \mathrm{K}\ , y_n)$, that is to find the Boolean decision function

$$f(Y) = f(y_1, \mathrm{K}\ , y_n) : \{-1,\ 1\}^n \to \{-1,\ 1\},$$

which gives the most reliable reconstruction of z.

We'll consider a probability model of classifiers with following properties:

1. *a priory* probabilities be equal
$$\Pr\{z = 1\} = \Pr\{z = -1\} = 1/2;$$
2. for each classifier the probability of correct classification does not depend on belonging of classified object and is at least
$$\Pr\{y_i = 1 \mid z = 1\} = \Pr\{y_i = -1 \mid z = -1\} = p_i \geq 1/2;$$
3. the classifiers are statistically independent.

It's well known that the decision, minimizing probability of error in this case, is the threshold function $f(Y) = \mathrm{sgn}(w_1 y_1 + \mathrm{K} + w_n y_n)$, where $w_i = \log \dfrac{p_i}{1 - p_i}$, $i = 1, \mathrm{K}\ , n$. The baricentric vector

$$W^{\beta} = \frac{1}{2} \sum_{Y \in \{-1,1\}^n} Y \Pr(Y \mid z = 1) - \frac{1}{2} \sum_{Y \in \{-1,1\}^n} Y \Pr(Y \mid z = -1) = (2p_1 - 1, \mathrm{K}\ , 2p_n - 1).$$

The probability of error for baricentric function is not exceed $\exp(-(W^{\beta})^2 / 2)$ [6, 7].

5. Accelerated perceptron in commettee decision

The accelerated perceptron in learning mode will be

$$W_{k+1} = W_k + z_k Y_k,$$

in self-learning mode

$$W_{k+1} = W_k + \mathrm{sgn}(W_k, Y_k) Y_k,$$

and majority function $f(Y) = \mathrm{sgn}(y_1 + \mathrm{K} + y_n)$ may be used for initial decision.

Theorem 1 *In the case of learning, the sequence of threshold functions generated by accelerated percepron stabilizes at baricentricl function* $f(Y) = \mathrm{sgn}(W^{\beta}, Y)$ *with probability 1.*

In order to get an analogous result for the case of self-learning, we need to define two new concepts.

Definition 1 *The vector* $\mathrm{M} = \mathrm{M}(Y) = \sum_{Y \in \{-1,1\}^n} \Pr(Y) f(Y) Y$ *is called the moment of the function f(Y).*

Definition 2 *The function f(Y) is called stable if* $f(Y) = \mathrm{sgn}(\mathrm{M}(f), Y)$.

Theorem 2 *In the case of self-learning, the sequence of threshold functions generated by accelerated perceptron stabilizes at one of the stable function with probability 1.*

Theorem 3 *If* $p_i \geq \frac{1}{2} + c$, $c > 0$, $i = 1,\mathrm{K},n$, *then moments of all stable functions in the positive orthant are of the form* $(2p-1)W^{\beta} + E$, *where* $|E| < 4\sqrt{n}\exp(-2nc^4)$ *and p is a probability of correct decision.*

Theorem 4 *If* $p_i > \frac{1}{2}$, $i = 1,\mathrm{K},n$, *then for every* $\varepsilon > 0$ *there is such a constant* $c(\varepsilon) > 0$, *that for every initial weight vector W belonging to the positive orthant and such that* $|W| > c(\varepsilon)$, *all the weight vector, generated by the accelerated perceptron in case of self-learning will belong to the positive orthant with probability 1.*

According to the three last theorems the decision function of accelerated perceptron in self-learning mode tends to baricentric decision function provided that a number of classifiers is large enough and they are of good quality. This guarantee an improvement of the decision function quality during self-learning.

References

[1] Duda R.O., Hart P.E., Pattern Classification and Scene Analysis, Wiley, New York, 1973.

[2] Minsky M., Papert, Perceptrons, MIT Press, Cambridge, MA, 1969.

[3] Nilsson N.J., Learning Machines, McGraw-Hill, New York, 1965.

[4] Pierce W.J., Failure-Tolerant Computer Design, Academic Press, New York, 1965.

[5] Rosenblatt F., Principles of Neurodinamics: Perceptron and the Theory of Brain Mechanisms.

[6] Zuev Yu.A., Ivanov S.K., Learning and self-learning in weighted voting procedures, Zh. Vychisl. Mat. Mat. Fiz., 35(1) (1995) 104-121.

[7] Zuev Yu.A., Ivanov S.K., The voiting as a way to increase the decision reliability, Journal of Franklin Institute, 336(2) (1999) 361-378.