Paper

Effectiveness of active forgetting in machine learning applied to financial problems

Hirotaka Nakayama and Kengo Yoshii

Abstract — One of main features in financial investment problems is that the situation changes very often over time. Under this circumstance, in particular, it has been observed that additional learning plays an effective role. However, since the rule for classification becomes more and more complex with only additional learning, some appropriate forgetting is also necessary. It seems natural that many data are forgotten as the time elapses. On the other hand, it is expected more effective to forget unnecessary data actively. In this paper, several methods for active forgetting are suggested. The effectiveness of active forgetting is shown by examples in stock portfolio problems.

Keywords — pattern classification, potential method, additional learning, forgetting.

1. Introduction

In many practical problems, e.g., financial investment problems, the situation changes very often over time. In machine learning, therefore, decision rules are needed to adapt for such changeable situations. To this end, additional learning should be made on the basis of new data. One of the authors and his collaborators have reported the effectiveness of additional learning in several machine learning techniques: mathematical programming approach [1], potential method [2] and RBF networks [3–5].

On the other hand, since the rule for classification becomes more and more complex with only additional learning, some appropriate forgetting is also necessary. Although several trials of forgetting in machine learning have been also suggested, they are concernced in such a way that the degree of importance of data decreases over time [3–5]. We call the way of forgetting based only on the time elapse "passive forgetting". However, it seems more effective to forget data which give bad influences to the current judgment. We call this way of forgetting "obstacle data" actively "active forgetting". In this paper, the effectiveness of active forgetting will be proved through some examples in stock portfolio problems.

2. Potential method

To begin with, the potential method suggested by one of the authors *et al.* [2] is reviewed briefly. The idea of potential method is originated from the static electric theory. Another similar method is the restricted Coulomb energy (RCE) classifier by Cooper [6] and Reilly *et al.* [7]. RCE tries to increase the ability of classification by adjusting the radia of hyperspheres which approximate the region of influence of data.

Unlike RCE, however, the potential method adjusts "charge" associated with each data in order to increase the ability of generalization. Each hidden unit corresponds to each teacher's pattern $x_j (j = 1, \cdots, N)$, which has some amount of charge c_j in which the sign depends on which category it belongs. Letting $D(x, x_j)$ denote a distance between x and x_j, the output unit is connected to

$$z(x) = \mathrm{sgn} P(x),$$

where

$$P(x) = \sum_{j=1}^{N} \frac{c_j}{D(x, x_j)}.$$

Here, P is the well known potential function, which sign decided on which category a given test data belongs to.

Note that the potential method can classify each teacher's data x_j $(j = 1, \cdots, N)$ correctly without doing anything, because $P(x_j) = +\infty$ for $c_j > 0$ and $P(x_j) = -\infty$ for $c_j < 0$. This means that the potential method can make the perfect learning for given teachear's data without doing anything. If we use the potential method as it is, however, it yields several small isolated influence regions just like "islands" in many problems. Clearly, this phenomenon causes a poor generalization ability. Therefore, in oder to obtain as smooth a discriminant surface as possible, we adjust the charges of data. This is the learning of the potential method.

A way of learning in the potential method can be summarized as follows:

Step 0. At the beginning, all teacher's data have an equal amount of charge except for the difference in its sign (suppose that each data of the class $\mathscr{A}$ has a positive charge, while each data of the class $\mathscr{B}$ a negative charge).

Step 1. Consider the ith pattern x_i $(i = 1, \cdots, N)$. Examine whether it is categorized correctly or not on the basis of the sign of output of

$$\tilde{P}(x_i) = \sum_{j \neq i}^{N} \frac{c_j}{D(x_i, x_j)}.$$

If the pattern x_i is not categorized correctly, then add the index i to the set I_{error}. If I_{error} is empty, then stop the iteration. Otherwise go to the next step.

Step 2. Find the pattern x_p with the highest error, namely

$$|\tilde{P}(x_p)| = \max_{i \in I_{error}} |\tilde{P}(x_i)|.$$

Step 3. Find the pattern x_q in the other category than of x_p nearest to the pattern x_p. Change the charge c_j of pattern x_j $(j = 1, \cdots, N)$ in such a way that the potential at $x_m = (x_p + x_q)/2$ becomes zero. Namely, suppose that the new charge $c_j^{'}$ is given by

$$c_j^{'} = c_j \exp\left(\mp \tilde{P}(x_j)\gamma\right) \quad j = 1, \cdots, N, \tag{1}$$

where denoting $q_j = c_j / D(x_m, x_j)$, $(j = 1, \cdots, N)$, γ solves

$$q_1 \exp\left(-\tilde{P}(x_1)\gamma\right) + \cdots + q_{N'} \exp\left(-\tilde{P}(x_{N'})\gamma\right) + \\ + q_{N'+1} \exp\left(\tilde{P}(x_{N'+1})\gamma\right) + \cdots + q_N \exp\left(\tilde{P}(x_N)\gamma\right) = 0. \tag{2}$$

Here, $x_1, \cdots, x_{N'}$ have positive charges, while $x_{N'+1}, \cdots, x_N$ negative charges. The sign $\mp$ in Eq. (1) means that $c_j > 0$ takes "$-$" and $c_j < 0$ "+".
Replace the charge of each pattern by the new one given by Eq. (1), and go to Step 1.

Remark 1. In changing charges, we focus our attention on a data whose position has the highest potential in the opposite category. It is possible to consider all data whose positions have potentials with the opposite sign. In this event, the equation to be solved becomes a system of several nonlinear equations. Although the authors examined several methods for solving the system of nonlinear equations, any technique have some difficulties, say, being trapped in local minima, no convergence sometimes, time consuming and so on. Although the above method based on the Eq. (2) produces just an approximate solution to our modification problem of charges, it shows good performance in our experiences.

Remark 2. The potential method belongs to a class of kernel methods for machine learning in which the approximate function is given by

$$f(x) = \sum_{j=1}^{n} K_j(x, x_j) y_j,$$

where $y_j = 1$ for $x_j \in \mathscr{A}$ and $y_j = -1$ for $x_j \in \mathscr{B}$. In addition, the kernel $K_j(x, x_j)$ is a symmetric function that usually (but not always) satisfies the following properties [8]:

(i)	$K(x, x') \geq 0$	nonnegative,
(ii)	$K(x, x') = K(\|\|x - x'\|\|)$	radially symmetric,
(iii)	$K(x, x) = \max$	takes on its maximum when $x = x'$,
(iv)	$\lim_{t \to \infty} K(t) = 0$	monotonically decreasing with $t = \|\|x - x'\|\|$.

The potential methods uses the kernel $K(x, x') = \frac{c}{\|\|x - x'\|\|}$ $(c > 0)$. In this event, the above property (iii) should be interpreted in such a way that the kernel has an infinite maximum when $x = x'$. Although the infinity property is not desirable in many mathematical analysis, it has a positive meaning in pattern classification problems. For cases in which the kernel is infinite at a test pattern, the potential at the test pattern has the correct sign without any learning. The only problem is that the generalization ability without adjustment of "charge" is poor in general. Therefore, the learning in the potential method is to adjust "charge" in order to increase the generalization ability.

Remark 3. The potential method can be extended by using a generalized potential

$$P(x) = \sum_{j=1}^{n} \frac{c_j}{\{D(x, x_j)\}^r}.$$

As r becomes larger, the influence of the data nearest to the test pattern gets larger. In the case of $r \to \infty$, therefore, the potential method with a generalized potential becomes the same as the k-nearest neighbour method with $k = 1$.

3. Additional learning

We can show that the additional learning can be made easily by using the potential method. Let x_t be a data added newly to the existing teacher's data. The procedure of additional learning can be divided into 1) the case in which x_t is classified correctly by the present rule, and 2) the case in which x_t is misclassified by the present rule. The details are as follows:

Case 1. When the new data x_t is classified correctly by the present rule, find a data x_a closest to x_t but in the different category of x_t. In addition, find a data x_b closest to x_a but in the different category of x_a. Let x_{abm} be the middle point of x_a and x_b, i.e., $x_{abm} = (x_a + x_b)/2$. If the potential of x_{abm} has a different sign from that of x_t, then put the charge c_t on x_t in such a way that we have

$$P'(x_{abm}) := P(x_{abm}) + c_t / D(x_t, x_{abm}) = 0.$$

Namely, we put

$$c_t = -P(x_{abm}) \times D(x_t, x_{abm}).$$

However, if the potential of x_{abm} has the same sign as that of x_t, then we do not put any charge on x_t (i.e, $c_t = 0$). The purpose of consideration of the potential of x_{abm} is to check whether the discriminant surface can be made correctly by adding x_t. Also, by excluding unnecessary data from additional learning, the computation time can be made shortened.

Case 2. When the new data x_t is misclassified by the present rule, find a data x_a closest to x_t but in the different category from x_t. Let $x_{atm} = (x_t + x_a)/2$. Then put the charge c_t on x_t in such a way that we have

$$P'(x_{atm}) := P(x_{atm}) + c_t / D(x_t, x_{atm}) = 0.$$

Namely, we put

$$c_t = -P(x_{atm}) \times D(x_t, x_{atm}).$$

4. Forgetting

If we make only additional learning according as some new knowledge are added, the newly obtained rule becomes more and more complex. Clearly, this does not give us a good effect in generalization ability of the method. Rather, it seems that unnecessary (or, inappropriate) rule in the present situation should be excluded. Human beings seem to grow up in such an adaptive way. Therefore, we should introduce forgetting as well as additional learning in machine learning.

How to forget is a difficult problem in machine learning. Maybe, one way is to forget unimportant data. In this event, we have to consider the degree of importance of data. In the potential method, the degree of importance for each data is considered to be given by the value of kernel function $K_i(x,x_i) = \frac{c_i}{D(x,x_i)}$.

4.1. Passive forgetting

In many situations, it seems natural that the degree of importance of data reduces as the time passes. A method for forgetting may be given by

$$c'_f = c_f \exp(-\alpha t),$$

where t denotes the time elapsed, α – the coefficient of forgetting, c_f – the original charge, and c'_f – the charge after t-time passed. Additionaly, it is supposed that the data x_f is extracted from the set of teacher's data, if t is beyond a threshold (the forgetting period).

The above method for forgetting depends only on the time elapse. However, it seems more effective to forget more actively data which give bad influences to correct judgment. We call the way of forgetting depending on the time elapse "passive forgetting", whereas the one of forgetting data with bad influence actively "active forgetting". We shall discuss the way of active forgetting in more detail below.

4.2. Active forgetting

A key for active forgetting is to find data giving a bad influence to correct judgment. We call such data "obstacle data". One way for finding obstacle data is given as follows. Suppose that a test pattern x_t is misjudged by the potential method. Let I_F denote the set of data in the other category from x_t. Removing a data $x_i \in I_F$, judge the category of test data x_t on the basis of its potential. If the judgment is correct, the data x_i is considered an obstacle data. Find such an obstacle data by checking all data $x_i \in I_F$.

Several ways for forgetting obstacle data is possible. Two simple ways (methods) are discussed below.

Method 1. Constant rate of forgetting with respect to the distance. The importance of obstacle data (i.e., the value of kernel) is decreased by controlling only the charge regardless the distance between the obstacle data and the test pattern. Let c_f denote the charge of the obstacle data x_f. A modified charge c'_f is given, for example, by

$$c'_f = \alpha c_f.$$

Here, the rate of forgetting α takes a value from [0,1].

Method 2. Increasing rate of forgetting with respect to the distance. In many cases, as the distance between a data x_i and the test pattern x_t becomes smaller, the influence of the data x_i becomes larger. Therefore, it seems natural to increase the rate of forgetting as the distance between the obstacle data and the test pattern becomes smaller. In this event, the value of kernel is controlled directly. One example is given by

$$K'_i = \alpha\beta K_i,$$

where α takes a value from [0,1], and β is given by

$$\beta = \frac{2}{1+e^{-\theta D(x_i,x_t)}} - 1.$$

The parameter θ is determined mainly by experience.

5. Applications to stock portfolio problems

5.1. Single stock investment

Our problem is to judge whether a stock is to be purchased or not. Seven economic indices are taken into account. We have the data in the 119 periods in the past for which it is already known to be purchased or not. We made a test of discriminant ability of the potential method taking the first 50 data as the teacher's ones, and examined the ability of classification for the rest 69 data. Figure 1 compares the result without additional learning and the one with additional learning with/without forgetting. Flags represent misclassified data. It can be observed that the additional learning provides a good effect in classification, in particular, around the period of 80's.

5.2. Portfolio mix problems

Our problem here is to make a portfolio mix among 213 stocks in the market. As in the previous subsection, it is known in the past 50 periods (1987.1-1991.2) whether each stock is to be purchased or not. In this event, each stock is considered in terms of 10 economic indices. The return rate is given by

$$\text{return rate} = \frac{\text{the highest price during the anteceding 6 periods}}{\text{current price}} - 1.$$

We judge a stock to buy if the return rate is over a certain threshold. In the following simulation, we suppose this threshold is 0.2 (Fig. 2).

Table 1
Index of advantage over the market

	Potential method		RBF network		1-NN method	
	Top30% Getting	Free Getting	Top30% Getting	Free Getting	Top30% Getting	Free Getting
Initial learning only	1.49 (63.0)	1.46 (100.4)	1.59 (63.0)	1.23 (103.8)	0.85 (63.0)	1.30 (130.3)
Additional learning (without forgetting)	1.40 (63.0)	1.64 (49.0)	1.85 (63.0)	2.37 (37.3)	1.05 (63.0)	1.65 (57.7)
Additional learning (with passive forgetting only)	1.58 (63.0)	2.17 (25.9)	1.75 (63.0)	2.84 (15.4)	1.01 (63.0)	1.60 (52.6)
Additional learning (with active forgetting only)	5.39 (63.0)	13.04 (49.9)	1.84 (63.0)	4.09 23.5	— —	— —
Additional learning (with active & passive forgetting)	6.62 (63.0)	24.99 (46.8)	2.11 (63.0)	8.48 (10.9)	— —	— —

The average number of invested stocks is indicated with a bracket.

Table 2
Index of advantage over the market for various forgetting schedules
(Free Getting active and passive forgetting)

Forgetting rate α	r									
	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
1.00	1.91 (49.1)	2.56 (49.4)	2.87 (49.1)	3.77 (48.4)	6.10 (47.6)	7.28 (46.1)	9.51 (45.3)	14.90 (45.3)	16.29 (43.7)	21.12 (45.4)
0.70	2.66 (49.2)	3.20 (48.9)	3.77 (49.0)	4.65 (47.7)	6.29 (46.4)	9.61 (45.7)	11.10 (45.3)	12.81 (45.3)	17.09 (43.7)	18.78 (45.5)
0.50	7.28 (46.1)	6.88 (45.9)	8.83 (45.6)	9.61 (45.7)	9.34 (45.8)	9.81 (44.8)	12.56 (44.7)	13.79 (44.3)	14.22 (44.0)	18.57 (44.5)
0.30	6.24 (46.4)	6.58 (45.8)	8.41 (45.2)	10.46 (44.1)	12.94 (44.1)	13.79 (44.3)	14.37 (44.1)	18.20 (44.5)	16.01 (44.7)	21.17 (45.7)
0.10	15.84 (43.5)	16.66 (44.4)	15.33 (44.2)	17.11 (44.7)	16.87 (44.0)	18.57 (44.5)	18.34 (45.3)	22.54 (45.8)	20.13 (46.3)	21.72 (47.8)
0.09	17.25 (44.5)	16.86 (45.9)	19.15 (44.8)	20.96 (45.3)	19.83 (45.5)	21.57 (45.2)	19.69 (45.5)	23.38 (45.9)	19.36 (47.2)	22.26 (48.2)
0.07	17.43 (44.9)	17.85 (43.7)	16.87 (44.5)	18.22 (44.6)	21.87 (45.1)	22.60 (45.3)	22.21 (45.6)	23.37 (46.2)	23.48 (46.7)	20.78 (48.3)
0.05	23.30 (46.8)	24.13 (46.8)	21.05 (47.0)	22.24 (47.1)	20.17 (46.6)	20.84 (46.5)	21.75 (47.2)	24.99 (46.8)	23.10 (47.9)	19.41 (48.6)
0.03	20.78 (47.1)	20.30 (47.6)	20.24 (47.6)	21.31 (47.4)	20.19 (47.6)	22.41 (47.5)	21.26 (48.1)	21.59 (48.0)	19.54 (48.3)	21.34 (48.5)
0.01	22.52 (47.5)	21.43 (48.1)	19.91 (48.2)	20.04 (48.3)	19.03 (48.4)	19.42 (48.5)	19.35 (48.7)	21.09 (48.5)	21.36 (49.0)	20.89 (49.2)

The number of average invested stocks is indicated with a bracket.

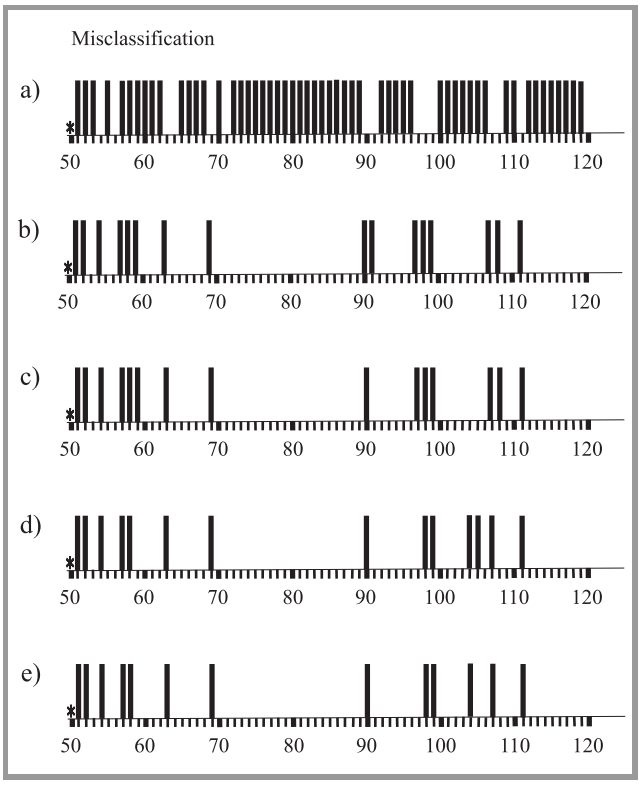

Fig. 1. Potential method with additional learning and forgetting: (a) initial learning, misclassified patterns 55; (b) only additional learning, misclassified patterns 16; (c) with passive forgetting, $r = -\lg 0.1/60$, misclassified patterns 15; (d) with active forgetting (method 1), $\alpha = 0.7$, misclassified patterns 14; (e) with active forgetting (method 2), $\alpha = 0.5$, misclassified patterns 13.

In the following, the way of Free Getting purchases only stocks which are judged to be purchased, while Top30% Getting the stocks of top 30% after sorting the stocks according to the degree of potential.

We made an examination of performance of portfolio mix by our method in the anteceding periods (1991.9-1996.3). Here, the index of advantage over the market I_A is defined by

$$I_A = \frac{(1+\alpha_1)(1+\alpha_2)\times\cdots\times(1+\alpha_T)}{(1+\beta_1)(1+\beta_2)\times\cdots\times(1+\beta_T)},$$

where α_i is the return rate of our portfolio mix at the ith period and β_i is the one of the market (usually called "index") at the ith period.

The initial learning was made for 50 periods between 1987.1 and 1991.2. The test with or without additional learning and forgetting is for 55 periods between 1991.9 and 1996.3.

The forgetting rates in cases with active forgetting only are $\alpha = 0.03$ and $r = 0.8$ for Top30% Getting, while $\alpha = 0.01$ and $r = 0.2$ for Free Getting. On the other hand, the forgetting rates in cases with active and passive forgetting $\alpha = 0.05$ and $r = 0.7$ for Top30% Getting, while $\alpha = 0.05$ and $r = 0.3$ for Free Getting. These values are the ones

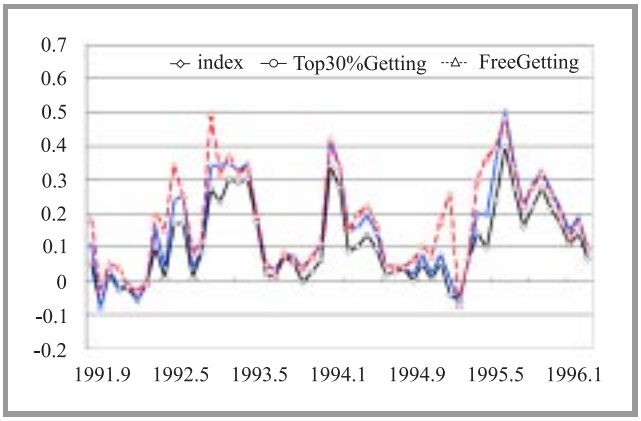

Fig. 2. Return rate by active and passive forgetting.

which provided the best result. Table 1 shows a comparison among potential method, RBF network and 1-*NN* method with various forgetting ways. The result for free getting with various forgetting schedules is shown in Table 2.

6. Concluding remarks

It has been observed that the effect of active forgetting is larger than that of passive forgetting. In general, the additional learning with forgetting provides a better performance than the mere additional learning. In the above example, however, the effect of appending forgetting to additional learning is not so remarkable in comparison with that of appending additional learning to the initial learning. In addition, the effectiveness of forgetting depends on its schedule. This implies that the forgetting is not so easy to use as the additional learning. It seems that human beings make forgetting in an effective way with almost optimal forgetting schedule on the basis of experience. Further examinations in practical problems are needed to find an optimal forgetting way in machine learning.

Acknowledgment

This work was supported by the Grant-in-Aid for Scientific Research of Education Ministry in Japan.

References

[1] H. Nakayama and N. Kagaku, "Pattern classification by linear goal programming and its extensions", *J. Glob. Opt.*, vol. 12, pp. 111–126, 1998.

[2] H. Nakayama and M. Yoshida, "Additional learning and forgetting by potential method for pattern classification", in *Proc. ICNN'97*, Houston, USA, 1997, pp. 1839–1844.

[3] H. Nakayama, "Growing learning mahines and their applications to portfolio problems", in *Proc. Int. ICSCC,* 1997.

[4] H. Nakayama, M. Yoshida, and S. Yanagiuchi, "Incremental learning for pattern classification", in *Proc. ICONIP'97*, Dunedin, New Zealand, 1997, pp. 498–501.

[5] H. Nakayama, S. Yanagiuchi, K. Furukawa, Y. Araki, S. Suzuki, and M. Nakata, "Additional learning and forgetting by RBF networks and its application to design of support structures in tunnel construction", in *Proc. Int. ICSC/IFAC Symp. Neural Comput. (NC'98)*, 1988, pp. 544–550.

[6] L. N. Cooper, "The hypersphere in pattern recognition", *Inform. Contr.*, vol. 5, pp. 324–346, 1962.

[7] D. L. Reilly, L. N. Cooper, and C. Elbaum, "A neural model for category learning", *Biol. Cybern.*, vol. 45, pp. 35–41, 1982.

[8] V. Cherskassky and F. Mulier, *Learning from Data: Concepts, Theory and Methods*. Wiley, 1998.

Hirotaka Nakayama
e-mail: nakayama@konan-u.ac.jp
Department of Information Science
and Systems Engineering
Konan University
8-9-1 Okamoto, Higashinada, Kobe 658, Japan

Kengo Yoshii
Department of Information Science
and Systems Engineering
Konan University
8-9-1 Okamoto, Higashinada, Kobe 658, Japan