Model Selection for Dynamic Processes

Shaomin Wu and Peter A. Flach

Department of Computer Science, University of Bristol,
Woodland Road, Bristol BS8 1UB, U.K
{shaomin,flach}@cs.bris.ac.uk

Abstract. In machine learning, ROC (Receiver Operating Characteristic) analysis is widely used in model selection when we consider both class distribution and misclassification costs that must be given at test time. In this paper we consider the case of a dynamic process, such that the class distributions are different in different time periods or states. The main problem is then to decide when to change models according to the different states of the generating process. In this paper we use a control chart to choose models for the process when misclassification costs are considered. Four strategies are considered and model selection approaches are discussed.

1. Introduction

In machine learning, ROC analysis for two classes measures the quality of models by studying the distribution of true positive rates and false positive rates of models. Both the class distribution and misclassification costs may be unknown during training time whereas they must be known at application time in order to select a suitable model. In practice, however, it may be difficult to know the exact class distribution which may change over time. In such cases, we need to know which point is a change point from one class distribution to another even when the class distributions in different periods may be known. Suppose instances $\{(\mathbf{X}_t, y_t), t=1,2,\ldots\}$ is a multivariate time series, where y_t is a binary class and $\mathbf{X}_t$ is a vector of independent features. From the ROC analysis point of view, we need to know the class distribution of y_t in order to choose a suitable model. In other words, we need to know where the change point from on state to another is.

The change-point detection has been discussed in [1,2,3]. A control chart, or cumulative count control chart (CCC-chart) can detect the change of class distributions that may be skewed in a process. This paper considers model selection by using CCC-chart.

This paper is organized as follows. Section 2 briefly reviews ROC analysis and CCC-chart. In Section 3, different situations from cost viewpoints will be taken into account and assumptions will be introduced. We distinguish four strategies for the different states of the process. Section 4 will give the costs for the four strategies and some analytical expressions for the average number of instances classified are derived. An example is given in section 5. Section 6 concludes.

2. ROC analysis and CCC-chart

ROC analysis [4, 5] studies the distributions of points (F,T) of models on a two-dimensional plane. Here, F stands for false positive rate (the ratio between the number of negative instances incorrectly classified and the total number of negative instances), and T stands for true positive rate (the ratio between the number of positive instances correctly classified and the total number of positive instances).

Assume that the relative frequency of negative instances in the test dataset is p. Assume that the cost for a correct classification is zero; the cost for classifying a positive instance to be a negative one is C_{pn} and the cost for classifying a negative instance to be a positive one is C_{np}. Then, the expected cost of applying model 1 with false positive rate and true positive rate (F_1,T_1) in the ROC space is $(1-p)(1-T_1)C_{pn}+pF_1C_{np}$. Similarly, the expected cost for model 2 is $(1-p)(1-T_2)C_{pn}+pF_2C_{np}$. Obviously, if $(1-p)(1-T_1)C_{pn}+pF_1C_{np}>(1-p)(1-T_2)C_{pn}+pF_2C_{np}$, then model 2 will be chosen. Otherwise, we shall choose model 1.

Assume labelled instances appear within a dynamic process one after another independently, and some candidate models can classify the instances into positives and negatives. An example would be a production line, where most items are manufactured correctly (positive) but some have production errors (negative). The number of positive instances until the next negative instance is observed is a geometric random variable. Let a process consist of two states S1 and S2 with relative frequencies of negative instances p1 and p2, respectively, where p1<p2. Let the probability of the event that the number of positive instances until a negative instance being observed is less than n_0 be α, or $P(n \le n_0)= \alpha$. Since n is a geometric random variable, we have $P(n \le n_0)= 1-(1-p_1)^{n_0}=a$ or $n_0=\log(1-\alpha)/\log(1-p_1)$ if the process is in S_1. Or if $n \ge n_0+1$, the process may be in S_1 with a probability $1-\alpha$ and n is called a type 1 signal (denoted as s_1). If $n \le n_0$, the process may have shifted to S_2 with a probability $1-\alpha$ and n is here called a type 2 signal (denoted as s_2). The approach here comes from CCC-chart methods[6, 7].

Because signal s_1 and s_2 show the state with a probability, they may be false ones. In order to confirm if a signal is true, an investigation may be carried out to check the true state of the process, which raise the different strategies in section 3. We assume that an investigation can recover the true state of the process. In what follows we make the following assumptions:

(1) Model 2 is more suitable for S_2 and model 1 is more suitable for S_1.

(2) When the process is in S_1, it may shift to S_2 with a probability π_{12}. When the process is in S_2, it may shift to another state with a probability π_{23}.

3. Four Different Strategies

One may decide whether a control chart will be used to monitor the process for different situations. We consider four possible strategies to decide when to switch between the two models.

(1) Strategy 1: In this strategy, no control chart will be used for the process. Because

the true state is not known, either model 1 or model 2 can be used throughout.

(2) Strategy 2: In this strategy, no control chart will be used. In order to know the exact state of the system, investigations for each instance are needed and two different models will be used according to the results of the investigations.

(3) Strategy 3: In this strategy, model 2 is used as soon as a signal s_2 appears. Although the signal s_2 may be a false one, no investigation on this signal will be carried out. In this strategy, the following two events may occur. Event A_1 — Before the process shifts to S_2, a signal s_2 appears when the process is in S_1, and then model 2 is used, and Event A_2 — in S_1, no signal s_2 appears. After the process shifts to S_2, a signal s_2 occurs when the process is in S_2, and then model 2 is used.

(4) Strategy 4: In this case, whenever a signal s_2 occurs, an investigation will be carried out to check the true state of the process. In this strategy, the following two events may occur. Event A_3 — before the process shifts to S_2, several s_2's occur and investigations are carried out. Model 2 is used until the process is confirmed to be in S_2 after a signal s_2 appears, and Event A_4 — in state 1, no signal s_2 appears. After the process shifts to S_2, a signal s_2 occurs in S_2 and an investigation is carried out and then model 2 is used.

4. Costs for the four strategies

Let $Q_{1(i)}$ (i=1,2) be the probability for signal s_i to appear when the process is in state S_1 and model 1 is being used, and $Q_{2(i)}$ (i=1,2) be the probability for a type i signal to appear when the process is in state S_2 and model 1 is still being used. Let $q_1=(1-p_1)(1-T_1)+p_1(1-F_1)$ and $q_2=(1-p_2)(1-T_1)+p_2(1-F_1)$, then we have $Q_{1(i)}=\sum_{j\in Z_i} q_1(1-q_1)^{j-1}(1-p_{12})^j$, $Q_{2(i)}=\sum_{j\in Z_i} q_2(1-q_2)^{j-1}(1-p_{23})^j$, where i=1,2

The probability of observing a transition of the process from state S_1 to state S_2 since the process starts is $Q_{1,2}=\sum_{j=1}^{\infty} p_{12}(1-q_1)^{j-1}(1-p_{12})^{j-1}$.

Recall that T_1 and F_1 are the true positive rate and the false positive rate of model 1, respectively, and T_2 and F_2 are the true positive rate and the false positive rate of model 2, respectively. Let the cost for investigating a signal be C_{in} and the cost for maintaining the CCC-chart be C_{chart}. Then, we can derive the following expressions for the expected cost for each of our four strategies.

Lemma 1. The expected cost for strategy 1 using model i is

$$c_{1i}=\left(\frac{1}{p_{12}}(1-p_1)+\frac{1}{p_{23}}(1-p_2)\right)(1-T_i)C_{pn}+\left(\frac{1}{p_{12}}p_1+\frac{1}{p_{23}}p_2\right)F_i C_{np}.$$

Lemma 2. The expected cost for strategy 2 is

$$c_2 = \left(\frac{1}{p_{12}}(1-p_1)(1-\mathrm{T}_1) + \frac{1}{p_{23}}(1-p_2)(1-\mathrm{T}_2)\right)C_{pn} + \left(\frac{1}{p_{12}}p_1\mathrm{F}_1 + \frac{1}{p_{23}}p_2\mathrm{F}_2\right)C_{np} + \left(\frac{1}{p_{12}} + \frac{1}{p_{23}}\right)C_{in}$$

Lemma 3. The expected cost for strategy 3 is

$$c_3 = P(A_1)\Bigg(E_1\big((1-p_1)(1-\mathrm{T}_1)C_{pn} + p_1\mathrm{F}_1C_{np}\big) + E_3\big((1-p_1)(1-\mathrm{T}_2)C_{pn} + p_1\mathrm{F}_2C_{np}\big)$$
$$+\frac{1}{p_{23}}\big((1-p_2)(1-\mathrm{T}_2)C_{pn} + p_2\mathrm{F}_2C_{np}\big)\Bigg) + P(A_2)\Bigg(E_2\big((1-p_1)(1-\mathrm{T}_1)C_{pn} + p_1\mathrm{F}_1C_{np}\big)$$
$$+E_4\big((1-p_2)(1-\mathrm{T}_1)C_{pn} + p_2\mathrm{F}_1C_{np}\big) + (\frac{1}{p_{23}} - E_4)\big((1-p_2)(1-\mathrm{T}_2)C_{pn} + p_2\mathrm{F}_2C_{np}\big)\Bigg) + C_{chart}$$

Lemma 4. The cost for strategy 4 is

$$c_4 = \frac{1}{p_{12}}\big((1-p_1)(1-\mathrm{T}_1)C_{pn} + p_1\mathrm{F}_1C_{np}\big) + E_4\big((1-p_2)(1-\mathrm{T}_1)C_{pn} + p_2\mathrm{F}_1C_{np}\big)$$
$$+(\frac{1}{p_{23}} - E_4)\big((1-p_2)(1-\mathrm{T}_2)C_{pn} + p_2\mathrm{F}_2C_{np}\big) + P(A_1)C_{in}E_5 + C_{in} + C_{chart}$$

where, $L_{1(i)} = \sum_{j\in Z_i} jq_1(1-q_1)^{j-1}(1-p_{12})^j$, $\quad L_{2(i)} = \sum_{j\in Z_i} jq_2(1-q_2)^{j-1}(1-p_{12})^j$, i=1,2.

$L_0 = \sum_{j=1}^{\infty}(j-1)p_{12}((1-q_1)(1-p_{12}))^{j-1}$, $p(A_1) = Q_{1(2)}/(1-Q_{1(1)})$, $p(A_2) = Q_{1,2}/(1-Q_{1(1)})$

$E_1 = (Q_{1(2)}L_{1(1)} + L_{1(2)}(1-Q_{1(1)}))/(1-Q_{1(1)})^2$, $\quad E_2 = (Q_{1,2}L_{1(1)} + L_0(1-Q_{1(1)}))/(1-Q_{1(1)})^2$,

$E_3 = 1 - p_{12}E_2P(A_2)/(p_{12}P(A_1)) - E_1$, $\quad E_4 = (Q_{2(2)}L_{2(1)} + L_{2(2)}(1-Q_{2(1)}))/(1-Q_{2(1)})^2$,

$E_5 = E_3/E_1$.

5. Example

Let p_1=0.002, p_2=0.008, T_1=0.995, T_2=0.990, F_1=0.004, F_2=0.002, C_{np}=1000, C_{np}=1 and α=0.05. It can be shown that we should use model 1 in S_1 and model 2 in S_2, respectively. When π_{12}=0.00002, π_{23}=0.00006, from Lemma 1, Lemma 2, Lemma 3 and Lemma 4, we can obtain

A. If C_{chart}=0 and C_{in}==0, then c_{11}=1265, c_{12}=1131, c_2=1081.5, c_3=1130.1 and c_4=1081.7, both strategy 2 and strategy 4 are the best cases;
B. If C_{chart}=0 and C_{in}=0.5, then c_{11}=1265, c_{12}=1131, c_2=34414, c_3=1130.1 and c_4=1111.6, strategy 4 is the best;
C. If C_{chart}==100 and C_{in}=0.5, then c_{11}=1265, c_{12}=1131, c_2=34414, c_3=1230.1 and c_4=1211.6, strategy 1 with model 2 used in both states S_1 and S_2 is the best.

To sum up, the data analyst can choose a strategy to minimize the cost. Say, when the cost for maintaining the CCC-chart is small or the cost for investigating the state of the system is small, strategy 3 or strategy 4 may be the best choice. In other words, maintaining the CCC-chart for the process is helpful in these cases.

6. Conclusions

When the class distributions of different states in the process are known and the change point of the states is not known, it is hard to apply different models for the different states. This paper combines both ROC analysis and CCC-charts to optimize the cost. Four different strategies have been considered and expressions for the expected costs for each of these strategies have been obtained. This aids the data analyst in deciding which strategy to choose under particular cost distributions.

Acknowledgement

This work is supported by the Esprit V project (IST-1999-11495) *Data Mining and Decision Support for Business Competitiveness: Solomon Virtual Enterprise*. Thanks are due to the anonymous reviewers for their comments and suggestions.

References

1. Wang, Y.: Change-point analysis via wavelets for indirect data. Statistica Sinica, 9 (1999) 103-118
2. Pignatiello, P., Samuelsabre, T., Estimation of the Change Point of a Normal Process Mean in SPC Applications. Journal of Quality Technology, 33(1) (2001), 82-95
3. Loader, C.: Change Point Estimation Using Nonparametric Regression, Ann. Statist., 24 (1996) 1667-1678
4. Ferri, C., Flach, P., Hernandez, J.: Learning decision trees using the area under the ROC curve. Nineteenth International Conference on Machine Learning. July 2002
5. Provost, F., Fawcett, T.: Robust classification for imprecise environments. Machine Learning, 42(3) (2001) 202-231
6. Calvin, T. W.: Quality control techniques for 'zero-defects'. IEEE Transactions on Components, Hybrid and Manufacturing Technology, CHMT-6(3), (1983) 323--328.
7. Goh, T. N.: A control chart for very high yield processes. Quality Assurance, 13(1) (1987) 18--22.

Appendix

It is easy to prove the following results. When the process is in state 1, the number of instances immediately before the process has shifted to state 2 is a geometric random variable with parameter π_{12}, and expectation $1/\pi_{12}$. The expected number of instances since the transition of the process from state S_1 to state S_2 until it shifts to another state, is $1/\pi_{23}$. Under event A_1, the expected number of instances since the start of the process until the appearance of the first signal s_2 in S_1 with model 1 being used is E_1. The probability of event A_1 is $p(A_1)$ and the probability of event A_2 is $p(A_2)$. Under event A_2, the expected number of instances since the start of the process until the time of the transition from state S_1 to state S_2 and no s_2 appearing during that time with model 1 being used is E_2. Under event A_1 or A_3, the expected number of instances since the time of the first appearance of signal s_2 until the transition from state S_1 to state S_2 is E_3. Under event A_2, the expected number of instances since the time of the transition from state S_1 to state S_2 until the appearance of the first signal s_2 in state 2 with model 1 being used is E_4. Under event A_3, the expected number of signal s_1 since the appearance of the first signal s_1 until the signal confirmed to be in state S_2 is E_5.

Proof of Lemma 1: The expected cost of applying model 1 in state 1 is $(1-p_1)(1-T_1)C_{pn}+p_1F_1C_{np}$. Similarly, the expected cost of applying model 1 in state 2 is $(1-p_2)(1-T_1)C_{pn}+p_2F_1C_{np}$. From the definition of strategy 1, and the above statement, we can obtain Lemma 1.

Proof of Lemma 2: If model 1 is being used in state 1 and model 2 is being used in state 2, the cost is $((1-p_1)(1-T_1)/\pi_{12}+(1-p_2)(1-T_2)/\pi_{23})C_{pn}+(p_1F_1/p_{12}+p_2F_2/p_{23})C_{np}$. In order to know the exact state of the system, investigations for each appeared instance are needed, the total cost for the investigation is $(1/p_{12}+1/p_{23})C_{in}$, then, we can get Lemma 2.

Proof of Lemma 3: For strategy 3,

(1) Under event A_1, the number of instances appearing before the appearance of the first signal s_2 in state 1 is E_1 and model 1 is being used during that time. The cost for this time period is $E_1((1-p_1)(1-T_1)C_{pn}+p_1F_1C_{np})$. The number of instances appearing since the appearance of the first signal s_2 until the time of the transition from state 1 to state 2 is E_3, and model 2 is being used during this time. The cost for this time is $E_3((1-p_1)(1-T_2)C_{pn}+p_1F_2C_{np})$. The number of instances since the system has shifted from state 1 to state 2 is $1/\pi_{23}$, then, the cost for this time period is $((1-p_2)(1-T_2)C_{pn}+p_2F_2C_{np})/\pi_{23}$.

(2) Under event A_2, the number of instances appearing in state 1 since the start of the process until the time of the transition from state 1 to state 2 is E_2 with model1 being used during that time. The cost for this time period is $E_2((1-p_1)(1-T_1)C_{pn}+p_1F_1C_{np})$. The number of instances appearing since the time of the transition from state 1 to state 2 until the appearance of the first signal s_2 is E_4,. The cost for this time period is $E_4((1-p_2)(1-T_1)C_{pn}+p_2F_1C_{np})$. The number of instances since the appearance of the first signal s_2 is $1/\pi_{23}-E_4$, then, the cost for this time period is $(1/\pi_{23}-E_4)((1-p_2)(1-T_2)C_{pn}+p_2F_2C_{np})$

To sum the above results of (1) and (2), and consider the probability of event A_1 and A_2, we get Lemma 3.

Proof of Lemma 4: For strategy 4, from the start of the process until transition from state 1 to state 2, model 1 is used; Between transition from state 1 to state 2 and appearance of the first signal s_2, model 1 is used. The number of instances in this time period is E_4.

(1) In state 2, after the first signal s_2 appears, model 2 is used; the number of the instances in this time period is $1/\pi_{23}-E_4$.

(2) In state 1, the number of investigations on signal s_2 when event A_3 and A_4 occur are $P(A_1)E_5$ and 0, in state 2, respectively. The number of investigations on signal s_2 when either event A_3 or A_4 occurs is 1.

Then, we can obtain Lemma 4