Using Artificial Neural Networks to detect Variance Change Point for Data Separation

Young Chul Han*, Kyong Joo Oh**, Tae Yoon Kim***

*Department of Information and Industrial Engineering, Yonsei University (pghyc@yonsei.ac.kr)

**Department of Information and Industrial Engineering, Yonsei University (johanoh@yonsei.ac.kr)

***Department of Statistics, Keimyung University (tykim@kmu.ac.kr)

Abstract

In this article, it will be shown that a nonparametric and data-adaptive approach to the variance change point (VCP) detection problem is possible by formulating it as a pattern classification problem. Technical aspects of the VCP detector are discussed, which include its training strategy and selection of proper classification tool.

Keywords : Variance change point detection; Nonparametric and data-adaptive method; Pattern Classification; Artificial neural networks; Discriminant analysis

1. Introduction

The problem of testing the parameter constancy of a time series has received considerable attention from researchers; see, for example, Bagshaw and Johnson (1997), Picard (1985), Kramer et al. (1998), Tang and MacNeil (1993) and papers therein. In particular, the volatility constancy problem has been an active research area since it was realized that many time series could be understood and predicted much better by analyzing their volatility. For examples of these kinds, see Engel (1982) for economic time series models with conditional heteroscedastic errors, Baufays and Rasson (1985) for modeling stock returns, and Kim et al. (2004) for an early warning system against economic crisis. In this context, detection of variance change point (VCP) has attracted much attention these days. Though the VCP detectors could be easily constructed by a single test statistic such as variance ratio type estimator, usually it is not easy to establish a nonparametric or data-adaptive detector since theoretical verification of such detector against various underlying stochastic processes is particularly difficult. In this article, it will be demonstrated that many technical breakthroughs of the VCP detection problem are possible by formulating it as a pattern classification problem. Formally, the VCP problem is stated as follows:

Suppose that $X_1, X_2, \ldots, X_n$ *is a sequence of time series data, and let*

$$\frac{Var(X_{i+1})}{Var(X_i)} = \alpha_{i+1} \qquad \text{for } i = 1, 2, \ldots, n-1 \quad (1)$$

Then, a sequence $X_1, X_2, \ldots, X_n$ *is said to experience a volatility change of size* α_ν *at VCP* ν *if* $\alpha_\nu \neq 1$.

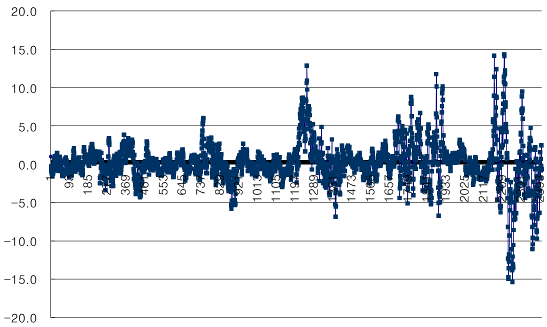
See Figure 1(a) which depicts time series data experiencing various volatility changes periodically. A typical approach to VCP detection is to employ a proper single test statistic for ν (e.g., variance ratio type estimator) and then provide theoretical justification for its practical applications. Through this approach, however, it is not easy to find a nonparametric or data-adaptive statistical solution since theoretical verifications of a single test statistic against various time series models including nonlinear or non-stationary autoregressive processes are particularly difficult. For example, CUSUM (cumulative sum) of squares defined by

$$V_{k,n} = \left| \frac{\sum_{j=1}^{k} X_j^2}{\sum_{j=1}^{n} X_j^2} - \frac{k}{n} \right| \quad \text{for} \quad 1 \le k \le n \quad (2)$$

has been investigated as a proper test statistic for ν since a sudden shape change of $V_{k,n}$ may indicate an existence of variance change (see Figure 1(b)).

However, its theoretical justification often turns out quite tough due to the fractional structure of $V_{k,n}$ entangled with the assumed time series model. For example, Lee and Park (2001) consider an infinite order moving average model to verify that $V_{k,n}$ converges to a standard Brownian bridge. These kinds of justifications for $V_{k,n}$ are still needed for other popular time series models (e.g., nonlinear or non-stationary AR) for its wide practical applications.

(a) X_i



(b) $V_{k,n}$

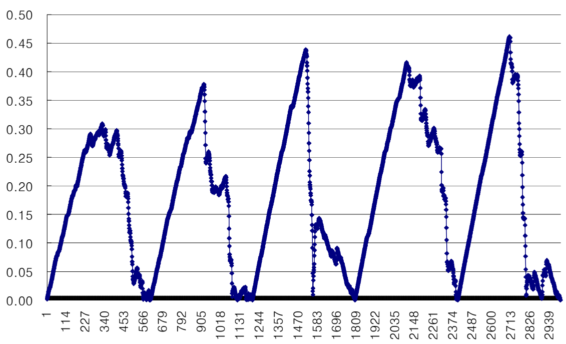


Figure 1. Time plots plots for X_i and its $V_{k,n}$ when $\alpha_{241}=4, \alpha_{721}=4, \alpha_{1201}=9, \alpha_{1681}=16, \alpha_{2161}=25$

The main goal of this paper is to show that a nonparametric or data-adaptive detector could be obtained if the VCP detection problem is formulated as a pattern classification problem. This is based on the observation that volatility change always causes an abrupt change at or around ν (VCP) in movements of time series or related statistics such as $V_{k,n}$ so that one can easily distinguish "a brief transition period containing ν" from an other stable period (see Figure 1(a)). Of course, successful formulation presupposes a proper training dataset containing a VCP and usually it is not difficult to obtain such dataset from the past of the given time series.

This paper consists of four sections. Section 2 is devoted to the formulation of VCP detection as pattern classification. Section 3 discusses technical aspects of the detector as a pattern classifier including training strategy and selection of appropriate classifier. Some simulation results are given in that section to examine the discussed technical aspects of the detector. Section 4 contains the concluding remarks.

2. Formulation of VCP Detection as Pattern Classification

This section begins with the definition of VCP training dataset. Let $X_1, X_2, \ldots, X_n$ and $X_{1(t)}, X_{2(t)}, \ldots, X_{m(t)}$ be two datasets of size n and m with volatility change of size α_ν and $\alpha_{\nu(t)}$ respectively, where $\nu \in \{1, 2, \ldots, n\}$ and $\nu(t) \in \{1(t), 2(t), \ldots, m(t)\}$. A numerical quantity subscripted by (t) simply denotes that it belongs to the training dataset. Then, we have:

Definition. $X_{1(t)}, X_{2(t)}, \ldots, X_{m(t)}$ *is a VCP training dataset of* $X_1, X_2, \ldots, X_n$ *if the underlying probability structure of* $X_{1(t)}, X_{2(t)}, \ldots, X_{m(t)}$ *is the same as that of* $X_1, X_2, \ldots, X_n$ *except for a possible different size of volatility change at VCP.*

Throughout this section, we assume that a VCP training dataset is given and simply refer to it as the training dataset. Once the training dataset is given, the next step for formulating VCP detection as a pattern classification problem is to establish template periods for training, i.e. to partition $X_{1(t)}, X_{2(t)}, \ldots, X_{m(t)}$ into three periods or patterns: Low volatility period (LP), brief transition one containing $\nu(t)$ (TP), and high volatility one (HP). This partitioning step is particularly critical for successful formulation, which is one of the main issues in Section 3 below. For now, let us assume an appropriate partition (LP, TP, HP) of the training period is found. Let C_n be a (pattern) classifier to be trained on the partition (LP, TP, HP). Notice that C_n signals volatility change whenever it reaches TP. If the training is successful (i.e., the trained C_n classifies the three periods efficiently on the training dataset), then it could be regarded as an empirical verification of C_n against the training dataset. *In other words, training* C_n *against the training dataset corresponds to verifying test statistic* V_n *against the assumed time series model.* As stated in the definition of VCP training dataset, it is noteworthy that its existence does not require any specific underlying time series models.

That C_n is equipped with such empirical verification benefits it with many desirable statistical properties. First, C_n is nonparametric since it is not necessary to assume any particular parametric form for the underlying structure of $X_1, X_2, \ldots, X_n$ for its application (recall that $X_1, X_2, \ldots, X_n$ is usually limited to simple stationary linear time series models for practical applications of various V_n's). Second, C_n is data-adaptive since it easily adapts itself through an empirical training process. Further, various transformations of training data could be employed as input variables for efficient training. The discussions above lead to the following claim:

Claim: *Variance change point detection of* $X_1, X_2, \ldots, X_n$ *could be formulated as a nonparametric and data-adaptive pattern classification problem if the training on its training data* $X_{1(t)}, X_{2(t)}, \ldots, X_{m(t)}$ *is done properly.*

3. Technical Issues

Given training data $X_{1(t)}, X_{2(t)}, \ldots, X_{m(t)}$ of size m, there are two major technical issues regarding the training of C_n: (i) how to construct a partition (LP, HP, TP), and (ii) what type of classifier is to be employed as a detector. For the first issue, TP is defined as

$$\text{TP}=\{(\nu-j)(t): j=p, p-1, \ldots, -p\} \qquad (3)$$

where $j=0$ implies a VCP ($\nu(t)$) experiencing $\alpha_{\nu(t)} = Var(X_{\nu(t)})/Var(X_{(\nu-1)(t)})$. Note that the relative length of TP of (3) to the entire training period is given by

$$l_{tp} = \frac{2p+1}{m} \qquad (4)$$

It is easy to see that small l_{tp} causes training inefficiency for TP while large l_{tp} difficulty in distinguishing TP from LP and HP (recall that TP contains small segments both from LP and HP simultaneously). For $\alpha_{\nu(t)}$, one may easily expect that a large $\alpha_{\nu(t)}$ would facilitate the classification task among HP, LP, and TP. Therefore, l_{tp} and $\alpha_{\nu(t)}$ are two important parameters to be determined for the optimal construction of a partition. For the second issue, there are many classifiers available from statistical machine learning algorithms (1995, 1990). To select a proper classifier for the VCP detection problem, careful consideration must be given to the unique feature of TP (i.e., TP is always characterized by its briefness and unusualness) since common classifiers tend to overlook TP as outliers, frequently in a standard training situation. Hence, either a chosen classifier could be adjusted to overfit to such a seemingly abnormal TP segment or a classifier having an overfitting tendency should be chosen. In this regard, an artificial neural network (ANN) might be a desirable tool since its overfitting tendency is quite well known for its technical drawback to overcome (1998). In addition, it is known that an ANN can cover a greater range of complexity problems including nonlinearity and non-stationarity, and is superior in its generality and practical ease in implementation due to its powerful and flexible capability (1995). In fact, Kim et al. (2004) noted the usefulness of an ANN under a similar situation to the VCP detection problem.

Some simulations are performed in order to examine these technical issues. In this simulation study, the following standard autoregressive process is considered.

$$X_i = \rho X_{i-1} + \sqrt{1-\rho^2}\, a_i, \quad i = 1, 2, \ldots \qquad (5)$$

where $\rho = 0.9$ and a_i is an identically and independently distributed (i.i.d.) normal error process with mean 0 and variance σ_i^2. The starting value for (5) is randomly chosen but to diminish the effect of the starting value, the first 50 values from (5) are burned or abandoned. Note that X_i is then normally distributed with mean 0 and variance σ_i^2 and that σ_i^2 makes a jump at $i=\nu$ with ratio α_ν. In fact, various values of ρ are tried ($-1<\rho<1$) in simulation but only $\rho = 0.9$ is presented here since other values produce quite similar results. For investigation of the first issue, training and test datasets with equal size ($n=m=480$) are generated. More precisely, training data $X_{1(t)}, X_{2(t)}, \ldots, X_{480(t)}$ are generated with $\alpha_{\nu(t)} = 4, 25$ (i.e. $\sigma_{\nu(t)}^2$ jumps either from 1 to 4 or 1 to 25) at $\nu(t) = 241$ and test data $X_1, X_2, \ldots, X_{480}$ with $\alpha_\nu = 4, 9, 16, 25$ at $\nu = 241$. Since each value of variance ratio ($=\alpha$) corresponds to one simulated dataset, a simulated dataset is denoted by it hereafter. Note that there are four test datasets ($\alpha_\nu = 4, 9, 16, 25$) for each training dataset ($\alpha_{\nu(t)} = 4, 25$). For each given training set TP defined at (3) is constructed with $\nu(t) = 241$ and its relative length $l_{tp} = 3\%, 5\%, 10\%, 15\%$. Recall that TP is defined such that it contains two heterogeneous groups of data having unequal variances simultaneously. Various input variables based on X_t are employed to extract information in classifying LP, TP and HP. The output variable for the detector takes the value of 1, 2, and 3, each representing LP, TP and HP, respectively, and signals volatility change whenever it reaches 2. Table 1 provides a list of input variables for the data in Figure 1. Some of their time plots are given in Figure 2. Now, an ANN whose architecture is given by Figure 3 is trained on each training data and then tested against various testing datasets. Here backpropagation neural network (BPN) is employed with $7 \times 7 \times 1$ multi-layer perceptron, i.e., input layer of 7 nodes, hidden layer of 7 nodes, and output layer of 1 node. As an activation function, the logistic function is used with learning rate, momentum and initial weights given by 0.1, 0.1 and 0.3, respectively. Note that each ANN trained on $\alpha_{\nu(t)} = 4$ and 25 separately is tested against test datasets in the order of $\alpha_\nu = 4, 9, 16, 25$.

Table 1. List of input variables

Name	Numerical formula	Description
DIFF	$d_t = X_t - X_{t-1}$	daily difference
MA(5), MA(20), MA(60)	$\mathrm{MA}(m) = \bar{d}_{m,t} = \frac{1}{m} \sum_{i=t-(m-1)}^{t} d_i$	m-day moving average
MV(5), MV(20), MV(60)	$\mathrm{MV}(m) = \frac{1}{m} \sum_{i=t-(m-1)}^{t} (d_i - \bar{d}_{m,t})^2$	m-day moving variance

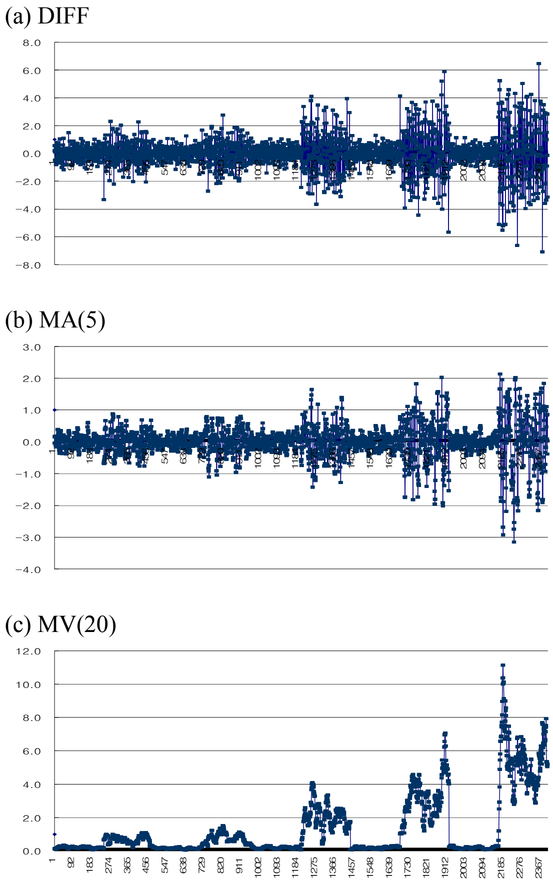


Figure 2. Time plots of some input variables

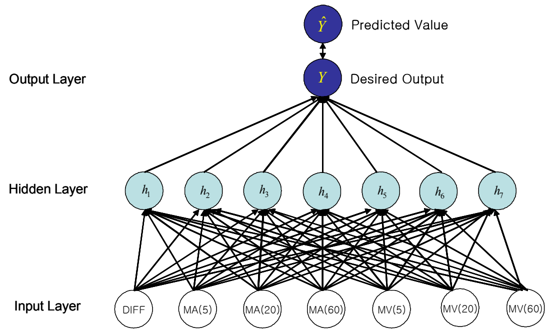


Figure 3. Architecture of the employed ANN

In Figure 4 (a) – (d) [Figure 5 (a) – (d)], training results for $\alpha_{v(t)}$ = 25 [$\alpha_{v(t)}$ = 4] are given first, and then four testing results are given in the order of $\alpha_v = 4, 9, 16, 25$ for each l_{tp} = 3%, 5%, 10%, 15% (recall that v and $v(t)$ are fixed as 241 for simulated datasets). Note that 480 data points are included in each result. In these figures, the desired output values are displayed together with their predicted output values for training and testing evaluation purpose. Careful examination of these figures reveals the following: (i) For each l_{tp}, C_n is trained better with $\alpha_{v(t)}$ = 25 than with $\alpha_{v(t)}$ = 4 , i.e., C_n with $\alpha_{v(t)} = 4$ has difficulty in training on LP and HP. (ii) C_n trained with $\alpha_{v(t)} = 4$ yields a better testing result than one trained with $\alpha_{v(t)} = 25$ in terms of correct classification hits. Particularly, C_n trained with $\alpha_{v(t)}$ = 25 has obvious problems on testing data with $\alpha_v = 4$. (iii) l_{tp} = 5% yields the best testing result with $\alpha_{v(t)} = 4$ while $l_{tp} = 3\%$ the best one with $\alpha_{v(t)}$ = 25 .

Each point mentioned above provides quite useful information in constructing a template partition (LP, TP, HP) of the given training period. First, a VCP detector is trained efficiently with large $\alpha_{v(t)}$. It is consistent with our intuition that VCP training data with dramatic change of variance would be easy to train. Second, a VCP detector trained with small $\alpha_{v(t)}$ tends to work better. This is due to the fact that small $\alpha_{v(t)}$ may much improve the sensitivity of detector. Third, an optimal l_{tp} is related to $\alpha_{v(t)}$ in the way that a small $\alpha_{v(t)}$ favors a rather long l_{tp}. It suggests that training inefficiency due to small $\alpha_{v(t)}$ could be improved by a rather long l_{tp}. From these, one may conclude that a good VCP detector would be obtained if it is trained with small $\alpha_{v(t)}$ and training inefficiency due to such small $\alpha_{v(t)}$ may be improved by selecting a rather long l_{tp}.

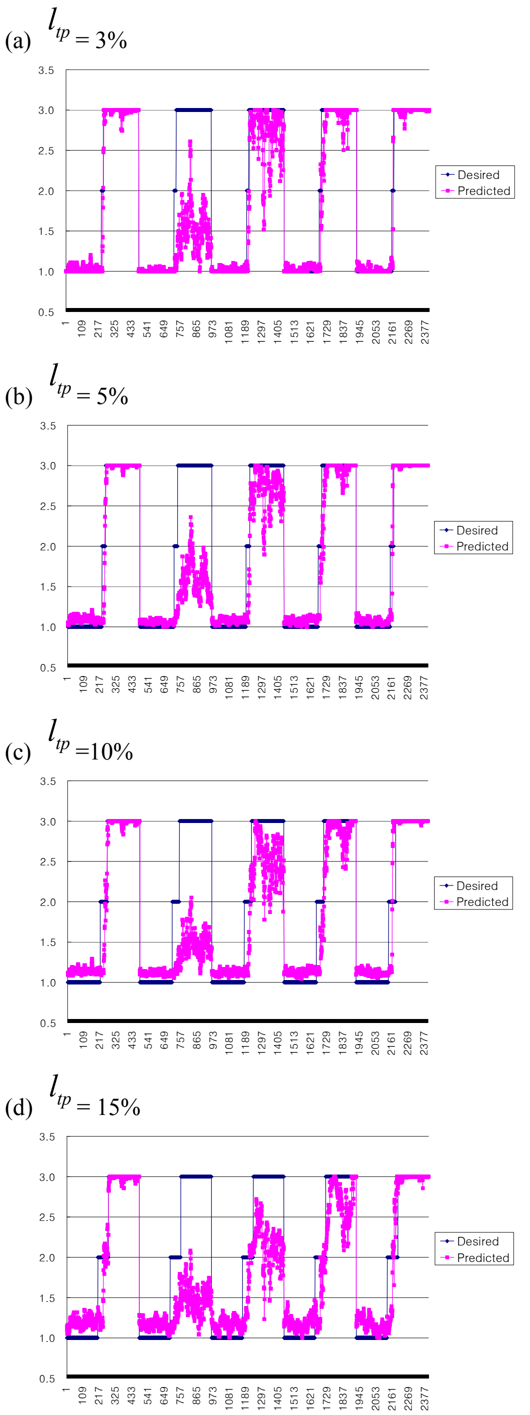


Figure 4. ANN for $\alpha_{v(t)} = 25$ and $\alpha_v = 4, 9, 16, 25$

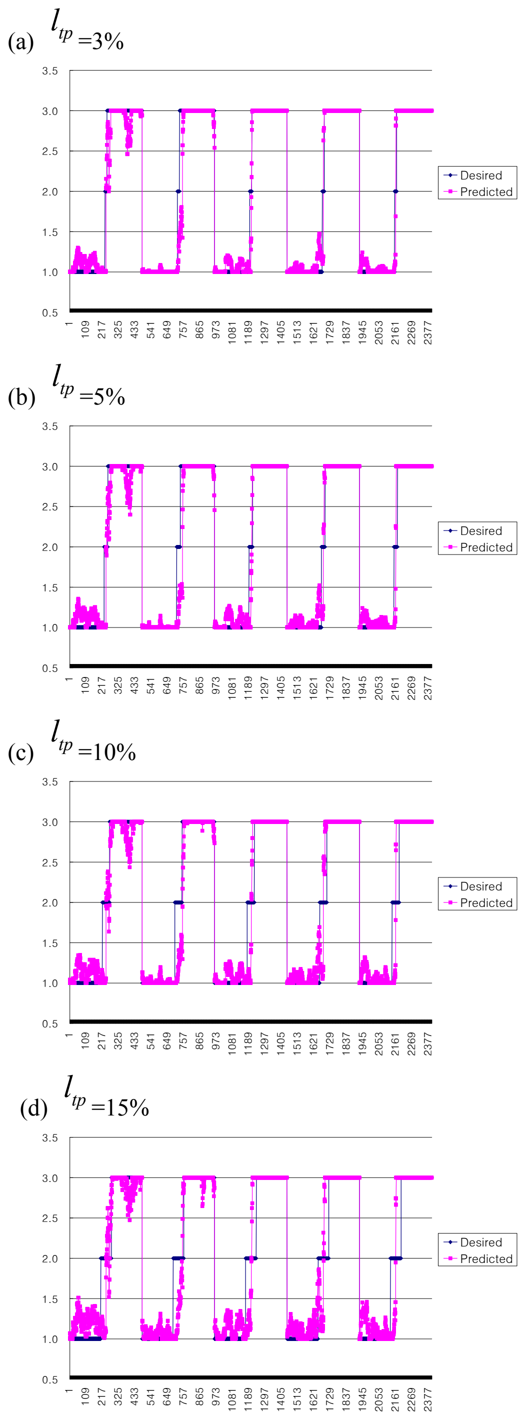


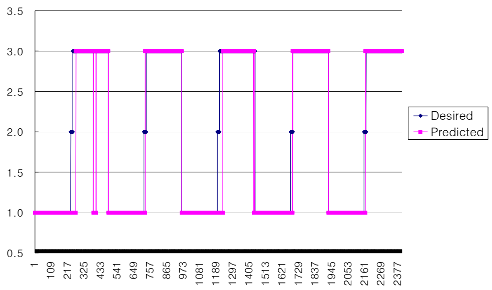
Figure 5. ANN for $\alpha_{v(t)} = 4$ and $\alpha_v = 4, 9, 16, 25$

As mentioned earlier, an ANN is employed here as a classification tool since ANN's major drawback, overfitting (1995, 2001), might work positively for a VCP detector. Indeed, the overfitting tendency of an ANN could be effective for concentrating on a very brief and abnormal segment of data (i.e., TP is usually a rare and abnormal segment). To support these points, an ANN is compared with multivariate discriminant analysis (MDA), a well-known parametric classifier. See Hair et al. (1995) or Mclachlan (1992) for detailed discussions of MDA.

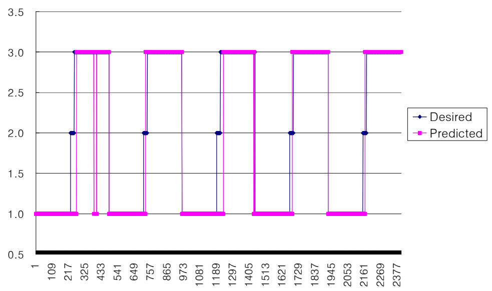
According to the recommendation from early simulation results, a simulated data with $\alpha_{v(t)} = 4$

and $\alpha_v = 4, 9, 16, 25$ is used. Classification results of MDA are given in Figure 6 (a) – (d) and show a very poor performance except for $l_{tp} = 15\%$. In fact, no correct detection of VCP is recorded for $l_{tp} = 3\%, 5\%, 10\%$, and a rather long $l_{tp} = 15\%$ reports some reasonable performance. Comparing these with the ANN performances given in Figure 5 clearly supports the ANNs superior capability in concentrating on brief and abnormal segments of data, which is a quite desirable strength as a VCP detector.

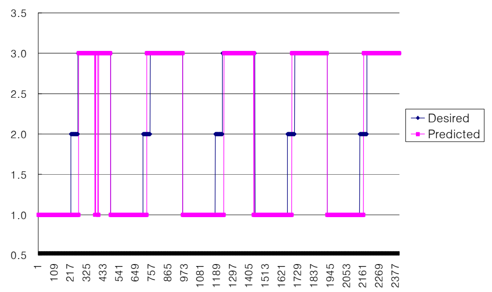
(a) $l_{tp} = 3\%$



(b) $l_{tp} = 5\%$



(c) $l_{tp} = 10\%$



(d) $l_{tp} = 15\%$

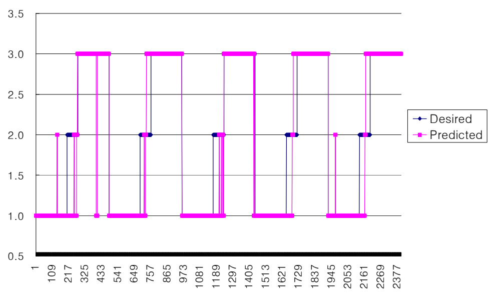


Figure 6. MDA for $\alpha_{v(t)} = 4$ and $\alpha_v = 4, 9, 16, 25$

4. Concluding Remarks

This article shows that the VCP detection problem could be formulated as a pattern classification problem. Such formulation deserves much attention since it could bring many technical breakthroughs to the VCP detection problem. Indeed, a nonparametric and data-adaptive approach to the VCP problem could be easily established whereas it is difficult for the classical approach based on single test statistics. Furthermore, one may utilize various features of time series movement at a VCP as input variables, which often greatly enhances the performance of a detector.

Discussions about training strategy and selection of proper classifier reveal various interesting aspects of the detector. An ANN is recommended as an appropriate tool since its well-known overfitting tendency could work positively for VCP detection. That a good VCP detector could be trained with a relatively small jump of variance (small $\alpha_{v(t)}$) over a rather long transition period (large l_{tp}) addresses basic issues for the selection of critical parameters in the VCP detection problem.

References

Bagshaw, M., & Johnson, R.A. (1977), Sequential procedures for detecting parameter changes in a time series model, *Journal of American Statistical Associations*, 72, 593-597.

Baufays, P., & Rasson, J.P. (1985), *Time Series Analysis: Theory and Practice*, North-Holland, New York, 119-127.

Bishop, C.M. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press, New York.

Engel, R.F. (1982), *Autoregressive conditional heteroscedasticity with estimates of variance of United Kingdom inflation*, Econometrica, 50(4), 987-1007.

Fukunaga, K. (1990), *Introduction to Statistical Pattern Recognition,* Morgan Kaufmann, New York.

Hair, J.F. Jr., Anderson, R.E., Tatham, R.L., & Balck, W.C. (1995), *Multivariate Data Analysis with Readings*, Prentice Hall, New York.

Kim, T.Y., Oh, K.J., Sohn, I., & Hwang, C. (2004), Usefulness of artificial neural networks for early warning system of economic crisis, *Expert Systems with Applications,* 26(4), 585-592.

Kim, T.Y., Oh, K.J., Kim, C., & Do, J.D. (2004), Artificial Neural Networks for Non-Stationary Time Series, *Neurocomputing*, 61, 439-447.

Kim, T.Y., Oh, K.J., & Kim, C. (2004), Daily economic condition indicator for early warning based on volatility, *submitted for publication*.

Kramer, W., Ploberger, W., & Alt, R. (1988), Testing for structural change in dynamic models, *Econometrica,* 56(6), 1355-1369.

Lee, S., & Park, S. (2001), The cusum squares test for scale changes in infinite order moving average process, *Scandinavian Journal of Statistics*, 28, 625-644.

Ling, C.X. (1995), Overfitting and generalization in learning discrete patterns, *Neurocomputing,* 8(3), 341-347.

McLachlan, G.J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York.

Nakashima, A., Hirabayashi, A., & Ogawa, H. (2001), Error correcting memorization learning for noisy training examples, *Neural Networks*, 14(1), 79-92.

Picard, D. (1985), Testing and estimating change points in time series, *Advances in Applied Probability,* 17, 841-867.

Prechelt, L. (1998), Automatic early stopping using cross validation: qualifying the criteria, *Neural Networks*, 11, 761-767.

Refenes, A.-P. (1995), *Neural Networks in the Capital Markets*, Wiley, New York.

Tang, S.M., & Macneil, I.B. (1993), The effect of serial correlation on tests for parameter change at unknown time, *Annals of Statistics*, 21, 552-575.