

Time-Based Detection of Changes to Multivariate Patterns

Jing Hu George Runger
Arizona State University

Abstract

Detection of changes to multivariate patterns is an important topic in a number of different domains. Modern data sets often include categorical and numerical data and potentially complex in-control regions. Given a flexible, robust decision rule for this environment that signals based on an individual observation vector, an important issue is how to extend the rule to incorporate time-based information. A decision rule can be learned to detect shifts through artificial data that transforms the problem to one of supervised learning. Then class probabilities ratios are derived from a relationship to likelihood ratios to form the basis for time-weighted updates of the monitoring scheme.

1 Introduction

Detection of changes to multivariate patterns is an important topic that has been widely studied in a number of different domains, such as multivariate statistical process control and data stream mining. In the field of statistical process control, Hotelling's T^2 [Hotelling, 1947] is widely used. However, the test statistic only incorporates the most recent observation. Performance can be improved from decision rules that incorporate time information (the history of the data). Consequently, time-weighted decision rules were derived from T^2 , such as a multivariate exponentially weighted moving average (MEWMA) [Lowry et al., 1992] and several multivariate cumulative sum (MCUSUM) [Testik and Runger, 2004] procedures.

The T^2 and related procedures were developed from multivariate normally distributed data. Although these procedures are robust to moderate departures from these assumptions, the use of Mahalanobis distance implies that numerical measurements are required, and that the control regions are elliptical in shape. In modern data sets with large numbers of both numerical and categorical variables, a more flexible robust method to monitor is needed. The focus here is on a class of monitors that is derived from artificial contrasts. [Hwang et al., 2004] proposed a simple, computationally feasible approach to detect changes to multivariate patterns. The control problem was transformed to supervised learning. Classification error rates were used to generate a control boundary for a specified type I error and each point was calculated to be inside or outside the boundary; a point outside generated a signal of a potential process shift. However, the decision only relied on the most recent observation and the benefits of time information would naturally seem to apply to modern data sets and flexible monitors. Consequently, time-weighted extensions for modern data sets (analogous to T^2 extensions) is the scope of this work and a general method that can be applied to a number of monitors is presented.

When only the most recent observation defines the decision, many equivalent statistics (monotonic transformations) can be monitored. Consequently, there is not a single choice for a time-weighted accumulation. In this work, an analogy of Hwang's work to a generalized likelihood ratio test is developed and estimates are used to define a control statistic that can be used to accumulate time information.

In the database and data mining literature, work has been done on processing data streams. However, only some of this work addresses the problems of change in a data stream. [Aggarwal, 2003] proposed a framework to diagnose changes based on velocity density estimation with only heuristics to find trends. [Ben-David et al., 2004] proposed a non-parametric statistic to detect changes based on the samples in a reference window and the current sliding window. However this work did not address high dimensionality nor a practical approach to decision limits.

2 Monitors from Artificial Contrasts and Supervised Learning

[Hwang et al., 2004] used in-control observations collected during normal operations and artificial data that were simulated to represent an alternative without a pattern. Artificial data were simulated from a multi-dimensional uniform distribution which encompassed the in-control observations. A class label is created. In-control observations and artificial data were used as the training data to build a classifier that formed a control procedure. [Hu et al., 2006] discussed alternative artificial data to highlight particular fault conditions. An example of a boundary learned from such a method is show in Figure 1.

The details follow. Suppose that $f_0(x)$ is an unknown probability density for the in-control data, and $f_1(x)$ is a specified reference density. Combine the original, in-control data set $\{x_1, x_2, \dots, x_{N_0}\}$ and a random sample of size N_1 drawn from $f_1(x)$. Also, create a response variable y with $y = 0$ and $y = 1$ for each sample from $f_0(x)$ and $f_1(x)$, respectively. A solution to this two-class classification problem defines a control region. Points with predicted $\hat{y} = 0$ are assigned to the “standard” or “on-target” class while $\hat{y} = 1$ assigns points into the “off-target” class. A constant (limit) associated with the classifier is used to adjust the relative magnitudes of the two types of errors (class 0 assigned to class 1, and vice versa). Furthermore, [Hu et al., 2005] provided a method to identify variables that contribute to a signal from such a monitor.

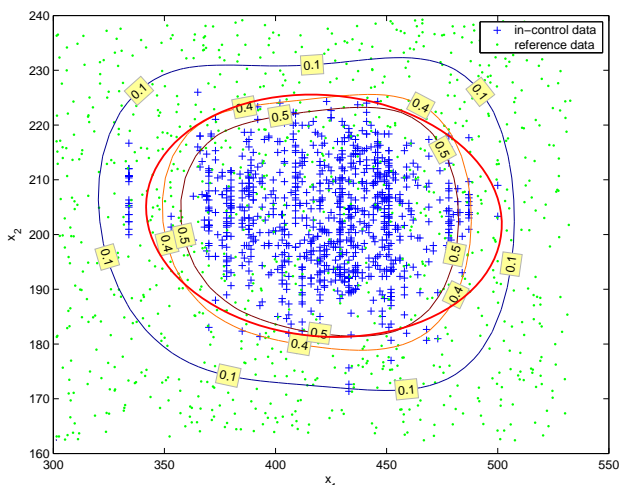


Figure 1: Example of control boundary learned from artificial contrasts versus Hotelling’s T^2 elliptical boundary.

3 Time-Weighting of Information

Section 2 provided a decision rule to monitor a process based only on the most recent observation. The manner in which information should be combined over time is still important. For example, based on a supervised learner with a control boundary, one might monitor cumulate the distance of each point to the control boundary over time. One might also cumulate the probability of an off-target condition.

Likelihood principles can be used as a guide to derive a general approach to incorporate time information. The general framework is the following:

$$H_0 : X_1, X_2, \dots, X_t \sim f_0(x)$$

$$H_1 : X_1, X_2, \dots, X_\tau \sim f_0(x), X_{\tau+1}, \dots, X_t \sim f_1(x)$$

for an unknown change time τ . The likelihood ratio is

$$L = \frac{\prod_{i=1}^{\tau} f_0(x_i) \prod_{i=\tau+1}^t f_1(x_i)}{\prod_{i=1}^t f_0(x_i)} = \frac{\prod_{i=\tau+1}^t f_1(x_i)}{\prod_{i=\tau+1}^t f_0(x_i)}$$

Because τ is unknown, a generalized likelihood ratio statistic maximizes the ratio of the likelihoods. The log of the generalized likelihood ratio is

$$l = \max_{\tau} \sum_{i=\tau+1}^t \ln \frac{f_1(x_i)}{f_0(x_i)}$$

Consequently, the control decision computes l (maximizing over all possible sample times for τ) and signals if the value exceeds a limit. For the case of univariate normal distributions, this computation is equivalent to a set of simpler recursive equations so that the control decision can be based on a Markov chain [Hawkins and Olwell, 1998]. However, in the multivariate case, even for a normal distribution, the result does not analogously simplify [Testik and Runger, 2004]. From the relationship between EWMA control charts and CUSUM control charts and likelihood ratios [Hawkins and Olwell, 1998], it is expected that a decision rule simpler than l can be effective. The change-point formulation provides the guidance to additively combine time information with the log likelihood ratio at time i

$$l_i = \ln \frac{f_1(x_i)}{f_0(x_i)} \quad (1)$$

Furthermore, any time-weighted control chart such as an EWMA or CUSUM can then be applied to the l_i 's.

3.1 Estimate of In-Control Probability

The following calculation from [Hastie et al., 2001] can be exploited to obtain an estimate of the statistic in Equation 1. According to Bayes' Theorem,

$$p(y = 1|x) = \frac{p(x|y = 1)p(y = 1)}{p(x|1)p(y = 1) + p(x|0)p(y = 0)} = \frac{f_1(x)N_1}{f_1(x)N_1 + f_0(x)N_0}$$

This algebraically leads to

$$\frac{f_1(x)}{f_0(x)} = \frac{p(y = 1|x) N_0}{p(y = 0|x) N_1} \quad (2)$$

If a supervised learner provides probability estimates (or an estimate of probability ratios), the density ratio can be estimated from direct substitution into Equation 2. For equal sample sizes, the log likelihood ratio is $l_i = \ln p_1(x_i) - \ln p_0(x_i)$. Some classification methods such as Regularized Least Square Classifier (RLSC) [Poggio and Smale, 2003] and decision-tree-based methods provide estimates of class probabilities. Also research has reviewed and modified existing classification methods for better estimation [Provost and Domingos, 2003, Niculescu-Mizil and Caruana, 2005].

Any classification methods that generates an estimate of probability ratios can be used with the statistic in Equation 1. In our experiment, a Random Forest (RF) is used as the supervised learning method [Breiman, 2001]. A random forest is an ensemble of tree predictors, each constructed from resampling the original observations. The forest prediction is the unweighted majority of class votes over all the trees. The test set error rates are monotonically decreasing and converge to a limit.

Thus, there is no over-fitting as the number of trees increase. The key to accuracy for RF is low correlation and bias. To keep bias low, trees are grown to maximum depth. To keep correlation low, the following randomization is used: 1) each tree is grown on a bootstrap sample of the training set; 2) at each node, m variables are selected at random out of the total number of variables p (m being much smaller than p). The default value for m is \sqrt{p} . But RF is not sensitive to the value of m over a wide range. RF has an internal mechanism to monitor generalization error. For every tree grown, about one third of the cases are out of the bootstrap (OOB) sample. The OOB samples can serve as a test set for the tree grown on the non-OOB data and provides unbiased estimates of the forest test set error. Another powerful feature of a tree-based method such as RF is that they naturally incorporate categorical as well as numerical variables. Results are also invariant to variable scales and robust to outliers among the predictors. The analysis scales well for large data sets.

3.2 Time-Based Detection Procedure

In Phase 1, in-control data are collected from the process. When no prior knowledge of the shift of the process is known, artificial data can be drawn from a multivariate uniform distribution. It could also be collected from previous known out-of-control situations.

In Phase 2, for each new sample x_t from the process, its in- and out-control probability \hat{p}_{0t} and \hat{p}_{1t} , respectively, are estimated by the classifier. A time-weighted control chart is updated with the ratio in Equation 1.

Any time-weighted control chart can be applied to l_t . For example, an exponentially weighted moving average (EWMA) is defined as $z_t = \lambda l_t + (1 - \lambda)z_{t-1}$ where $0 < \lambda \leq 1$ [Roberts, 1959]. The mean μ and the variance σ^2 of the likelihood ratio l_t can be estimated through the in-control observations. The control limits for the EWMA control chart (from the steady-state, approximate formula for the variance) are $CL = \mu \pm L\sigma\sqrt{\lambda/(2 - \lambda)}$.

4 Simulation Examples

Several experiments are presented. Each experiment is replicated 10 times and the average and the standard deviation of the average run length for on-target and off-target shifts are reported (ARL0 and ARL1, respectively). For each experiment, 500 replicates are conducted to evaluate the ARL0 and ARL1. The control limits use $L = 2.96$ and the smoothing parameter is $\lambda = 0.2$.

Bivariate normal data with independent and then correlated elements are considered. In both cases the in-control mean vector is $(0, 0)$ and the variables have unit variances. For the correlated case the correlation coefficient is 0.7. Table 1 shows the ARLs for an EWMA for the independent case and Table 2 shows the ARLs for EWMA, xbar, and individuals control charts based on the statistic in Equation 1 for the correlated cases. The summary statistics for the mean, variance of l_t and the calculated control limit are also shown in the tables. The sample sizes are equal and $N_0 = N_1 = 5000$.

Table 1: ARLs for an EWMA applied to the l_t statistic for independent, bivariate normal data

	mean	variance	control limit	ARL0	ARL1 at Shifted Mean			
					(1,0)	(2,0)	(1,1)	(2,2)
Average	-2.75	5.4	-0.46	317.52	33.07	5.36	12.14	2.89
Standard deviation	0.14	0.3	0.16	54.35	3.74	0.35	1.01	0.12

Table 2: ARLs for schemes applied to the l_t statistic for a dependent, bivariate normal data

	mean	variance	control limit	ARL0	ARL1 at Shifted Mean					
					(1,0)	(2,0)	(1,1)	(2,2)	(1,-1)	(2,-2)
EWMA										
Average	-3.24	4.66	-1.11	275.6	12.37	2.81	25.9	4.45	3.23	2
Stdev	0.13	0.27	0.16	41.22	0.82	0.06	2.12	0.18	0.16	0
Individual										
Average	-3.18	4.73	5.9	233.05	27.09	3.66	42.25	6.5	4.83	1.18
Stdev	0.05	0.19	0.51	44.91	4.87	0.52	6.21	0.77	0.76	0.05
Xbar										
Average	-3.2	4.75	-0.28	230.56	19.42	5.18	24.31	6.99	5.46	5
Stdev	0.13	0.28	0.16	4.1	1.44	0.05	0.59	0.35	0.13	0

A MEWMA control chart designed for normally distributed data with $ARL0 \approx 200$ and $\lambda = 0.2$ provides the following results: $ARL1 = 3.8$ and $ARL1 = 10.1$ at a Mahalanobis distance $\delta = 2$ and $\delta = 1$, respectively [Lowry et al., 1992]. The $ARL0$ for the method in the tables is greater, and the $ARL1$ is comparable for shifts with $\delta \geq 2$. Not surprisingly, the MEWMA procedure designed for the known normal distribution excels for the smaller shifts of $\delta = 1$. The objective of this research is not to compare to normal theory, but to evaluate the time-weighted information. The comparison of the EWMA, individuals, and xbar charts in Table 2 illustrates the advantages of the time-weighting applied to the artificial contrast method. As expected, the time-weighting becomes more effective as the magnitude of a shift decreases.

Table 3: ARLs for selected schemes applied to the l_t statistic for 10-dimensional, independent, normal data

	mean	variance	control limit	ARL0	5 vars shift 1σ	10 vars shift 1σ
EWMA						
Average	-4.62	4.90	-2.43	357.69	11.07	5.21
Stdev	0.08	0.18	0.11	73.02	1.30	0.26
Individual						
Average	-4.32	5.46	2.23	219.33	26.76	16.34
Stdev	0.09	0.25	0.28	39.61	9.59	5.24

Another experiment uses 10 dimensional, independent normal variables (Table 3). Here $ARL1$ is evaluated for two out-of-control situation: 5 out of the 10 variables shift 1σ in the mean and all the 10 variables shift 1σ in the mean. The sample sizes are equal and $N_0 = N_1 = 2000$. The results are also compared with those of the individual charts. In high dimensions, the magnitudes of shifts in the table are still quite small so weighting information over time excels.

5 Conclusions

We developed a methodology to incorporate time information into a control problem transformed to supervised learning. Whenever a learner provides class probability memberships the analysis here can be applied to generate a more sensitive control algorithm. Consequently, the transform along with the statistic to monitor makes for a conceptually simple procedure with a broad range of applicability. Data of any type can be blended in the analysis. Representative examples were

provided for a random forest learner and data with known distributions. The results illustrate the benefits of time information in the monitor.

References

- [Aggarwal, 2003] Aggarwal, C. (2003). A framework for diagnosing changes in evolving data streams. In *Proceedings of the ACM SIGMOD Conference*, San Diego, California.
- [Ben-David et al., 2004] Ben-David, S., Gehrke, J., and D., K. (2004). Detecting change in data streams. In *Proceedings of the 30th VLDB*, Toronto, Canada.
- [Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
- [Hastie et al., 2001] Hastie, T., Tibshirani, R., and Firedman, J. (2001). *Elements of Statistical Learning*. Springer, New York.
- [Hawkins and Olwell, 1998] Hawkins, D. M. and Olwell, D. H. (1998). *Cumulative Sum Charts and Charting for Quality Improvement*. Springer, New York.
- [Hotelling, 1947] Hotelling, H. (1947). *Techniques of Statistical Analysis*, chapter Multivariate quality control-illustrated by the air testing of sample bombsights, pages 111–184. McGraw-Hill, New York.
- [Hu et al., 2005] Hu, J., Runger, G., and Tuv, E. (2005). Contributors to a signal from an artificial contrast. In *Proceedings of the 2nd International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, Barcelona, Spain.
- [Hu et al., 2006] Hu, J., Runger, G., and Tuv, E. (2006). Tuned artificial contrasts to detect signals. *International Journal of Production Research*.
- [Hwang et al., 2004] Hwang, W., Runger, G., and Tuv, E. (2004). Multivariate statistical process control with artificial contrasts. *IIE Transactions*. To appear.
- [Lowry et al., 1992] Lowry, C., Woodall, W., Champ, C., and Rigdon, S. (1992). A multivariate exponentially weighted moving average chart. *Technometrics*, 34:46–53.
- [Niculescu-Mizil and Caruana, 2005] Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proc. 22nd International Conference on Machine Learning (ICML'05)*.
- [Poggio and Smale, 2003] Poggio, T. and Smale, S. (2003). The mathematics of learning: Dealing with data. *Notices of the AMS*.
- [Provost and Domingos, 2003] Provost, F. and Domingos, P. (2003). Tree induction for probability-based ranking. *Machine Learning*, 52:199–215.
- [Roberts, 1959] Roberts, S. W. (1959). Control chart test based on geometric moving averages. *Technometrics*, 42(1):97–102.
- [Testik and Runger, 2004] Testik, M. and Runger, G. (2004). Multivariate extensions to cumulative sum control charts. *Quality and Reliability Engineering International*.