Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

Learning patterns through artificial contrasts with application to process control

E. Tuv[1] & G. Runger[2]
[1]*Analysis & Control Technology, Intel Corporation, USA*
[2]*Department of Industrial Engineering, Arizona State University, USA*

Abstract

In manufacturing as well as other application areas there is a need to learn standard operating conditions in order to detect future changes or deviations. This is related to the even more general problem of detecting instances (cases, records) that are unusual compared to the bulk of the data (outliers). Examples of the problem are fault detection in chemical engineering and statistical process control. The outlier problem is ubiquitous.

If specific deviations are not *a priori* specified, this is a type of unsupervised learning problem. The focus here is on the important, practical case for modern data environments. That is, training data with multiple (usual many) variables of mixed types (without the expedient assumptions common in statistics of multivariate normality that rarely holds in practice).

An elegant technique is used to transform an unsupervised learning problem to a supervised one. This methodology uses an artificial reference distribution. For the focus here such a specific reference distribution requires appropriate properties. Then an effective, universal, and nonparametric supervised learner (a gradient boosting machine) is applied to the transformed problem. The results are then in a sense inverted to the original problem. Extensions are mentioned as well as additional insight that becomes available. An illustrative example is presented to justify the validity of this generic and general methodology.

1 Introduction

In manufacturing as well as other application areas there is a need to learn standard operating conditions in order to detect future changes or deviations. This is related to the even more general problem to detect instances (cases, records) that

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

are unusual compared to the bulk of the data (outliers). It is well-known that such data can excessively distort and influence continuous measurements and thereby affect the decisions obtained from an analysis.

If specific deviations are not *a priori* specified, this is a type of unsupervised learning problem. The focus here is on the important, practical case for modern data environments. That is, training data with multiple (usual many) variables of mixed types (without the expedient assumptions common in statistics of multivariate normality that rarely holds in practice). Examples of the problem are fault detection in chemical engineering and statistical process control (SPC). The outlier problem is ubiquitous. The motivation for our learning method is manufacturing so that we briefly summarize traditional approaches.

When measurements from multiple variables (attributes) are available the challenge is called multivariate SPC in the statistics literature. A traditional approach was proposed quite early based on Hotelling's T^2 statistic [1] as the metric to signal a change. The distance of $\mathbf{X}_t$, an observed $p - dimensional$ vector at time t, to a target vector μ is $(X_t - \mu)'\Sigma^{-1}(X_t - \mu)$, where Σ is the covariance matrix of X_t. The statistic, under the assumption of multivariate normality, follows a χ^2 distribution (with p degrees of freedom) when sufficient training data is available that estimation error in Σ can be ignored. The decision rule applied to a sequence of vectors (assumed independent over time) is the appropriate percentile from this distribution and this defines a control region that has the form of an ellipse in p dimensions. The decision based on a single data point is well known to be improved by filtering methods such as multivariate moving averages or multivariate likelihood ratio methods. However, all these alternatives are rooted in normal theory (and it inherent continuous measurements) and therefore result in elliptical-based boundary (although the decision rule is more complex).

The problem has also been addressed from principal components analysis (PCA). This has been used in chemical industry for a decade for multivariate process diagnosis and fault detection. Applications to process control were considered as least as early as by Jackson [2]. Recently the semiconductor industry has slowly adopted the methodology in manufacturing while more and more correlated process and quality parameters become available and more rigorous monitoring is required for 90 nm process on 300 mm wafers. In an attempt to reduce the dimensionality of the problem, a lower-dimensional subspace is the focus. The data is projected to a lower-dimensional subspace that maximizes the variability of the projected points [2]. Also, a measure of deviation from the subspace is commonly monitored [2]. The major axis of the subspace is aligned with the major axis of the data, that is, the eigenvector corresponding the greatest eigenvalue of the covariance matrix. But the implicit development from the covariance matrix assumes the data is characterized only by first and second moments and consequently the multivariate normal distribution is the foundation for the approach. However, many parametric data are not normally distributed such as n-channel and p-channel leakages on a transistor. Furthermore, in semiconductor manufacturing process, many characteristics are associated to non-numeric information such as left zone and right zone on a wafer. Then these kinds of parameters cannot be

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

used in the PCA method. Our approach below improves upon these limitations of PCA.

Detection is often the first step, but application typically require more details from an analysis. Successful fault detection leads to the question of fault diagnosis. A method to detect should consider the consequences: corrective action is required to adjust the process back to desirable operating conditions. With multiple variables the diagnosis can be a challenge and although the problem has been well studied there are extensive computational and performance limitations of the proposals. Runger *et al.* [3] proposed a computationally simple approach for the normal theory case. Outside of the comfort of the well-studied normal theory, a general strategy for diagnosis is important. The approach should be conceptually straightforward in order to facilitate its adoption, but comprehensive enough to incorporate non-normal variables of different data types. Such a method is integrated into our fault detection strategy.

In this paper, we propose a supervised classification approach to learn the pattern of standard operating conditions, and thereby detect deviations from the pattern. Our method does not require that the deviation be *a priori* specified. This is important in our manufacturing application in which there are many variables of mixed types and the deviations can exhibit various characteristics. We use an interesting approach that transforms the problem to supervised learning through an artificial contrast. Furthermore, we extend our methods to address the important question of fault diagnosis.

2 Supervised pattern learning

2.1 Transformation to a density estimation problem

In practice, the joint distribution of the variables of interest is unknown and rarely as well-behaved as a multivariate normal distribution. Instead, data describing the standard operation of the process is available and a verifiable method is needed to assign a future observation to the "fault" class with given confidence. In other words, we are looking for a probabilistic decision rule which will assign a new observation to the "fault" class of a given population. Thus we have binary classification problem. Motivation for the approach described in this section is from an elegant technique (considered to be "statistical folklore") to transform a density estimation problem to the one of function approximation [4].

Suppose $f(x)$ is a unknown probability density function to be estimated, and $f_0(x)$ is a specified reference density function (for example uniform). Combine the original data set $x_1, x_2, \ldots, x_N$ and a random sample of size N drawn from $f_0(x)$.

If we assign $y = 1$ to each sample point drawn from $f(x)$ and $y = 0$ for those drawn from $f_0(x)$, then

$$p(x) = E(y|x) = \frac{f(x)}{f(x) + f_0(x)} \tag{1}$$

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

Therefore we can solve the regression problem on the sample $(x_1, y_1), (x_2, y_2), \ldots, (x_{2N}, y_{2N})$ to obtain an estimate of the unknown density $f(x)$

$$\hat{f}(x) = f_0(x)\frac{\hat{p}(x)}{1-\hat{p}(x)} \tag{2}$$

For the binary classification problem considered in this work, a natural reference distribution is formed as a multivariate joint density of independently distributed X variables

$$f_0(x) = \prod_{i=1}^{k} f_i(x_i) \tag{3}$$

A sample from the independent joint density can be generated by randomly permuting values for each variable in the base data set (independently). Reference data could also be generated from uniformly distributed random values from the corresponding range of X-values for each numerical variable. For a categorical variable X with l_x levels one can assign nominal values with equal probabilities $1/l_x$. In any case, we still form the training sample (x_i, y_i) from the combined data.

2.2 Classification engine

There are several major advances in generic, supervised learning recently published in the statistical and machine learning literature that can be used to solve this problem: SVM (Support Vector Machine) [5], MART (Gradient Boosting Machine) [6, 7], RF (Random Forests) [8], and RLSC (a recent regularization algorithm) [9, 10]. Both RF and MART are tree-based algorithms, and therefore inherit all the powerful features of CART: natural handling of mixed data types and missing data, robustness to noise and outliers. In addition to these properties, MART and RF show significantly improved prediction accuracy.

In this work we use the gradient boosting machine MART as the classification engine. This approach approximates a response by an "additive" expansion of the form

$$F(\mathbf{x}) = \sum_{m=0}^{M} T(\mathbf{x}, \Theta_m) \tag{4}$$

where $T(\mathbf{x}, \Theta_m)$ ("base learner") is an small regression tree of fixed size (a "base learner"). We use trees with 10 terminal nodes. Also, $\Theta_m = \{R_{jm}, c_{jm}\}_{j=1}^{L}$ defines the tree's partition/prediction at m^{th} stage. More details for MART follow.

For a binary classification problem, logistic modelling generalizes

$$p(x) = \frac{e^{f(x)}}{1+e^{f(x)}} \tag{5}$$

with the *deviance* loss function

$$L(y, p(x)) = -I(y=1)log(p(x)) - I(y=-1)log(1-p(x)) \tag{6}$$

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

An estimate of $f(x)$ is obtained from a greedy stage-wise procedure. This is referred to "stage-wise" because of new term in the additive expansion is added at each iteration. At each step the procedure needs to solve

$$\hat{\Theta}_m = arg \min_{\Theta_m} \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + T(x_i, \Theta_m)) \tag{7}$$

where $\Theta_m = \{R_{jm}, c_{jm}\}_{j=1}^{L}$ define tree's partition/prediction at m^{th} stage.

Note that for binary classification with exponential loss this stage-wise procedure is equivalent [4] to the AdaBoost algorithm [11]. For more robust loss functions (such as deviance) approximations are needed to solve (7). Therefore, in general, MART uses a numerical optimization approach to solve (7). Specifically, it mimics a steepest descent algorithm applied to a specified target (loss) function with respect to a vector of parameters (values of an approximating function at N data points).

For a binary classification problem MART's forward "stage-wise" procedure can be summarized as follows (where ν is a learning rate, usually $\leqslant 0.1$):

1. Initialize $f_0(x) = 0$
2. For m=1 to M do:
   Set $p_m(x) = \frac{e^{f_m(x)}}{1+e^{f_m(x)}}$
   Compute $r_{im} = y_i - p_m(x),\ i = 1, \ldots, N$
   Fit a regression tree to the target r_m giving terminal regions $R_{jm},\ j = 1, 2, \ldots, J_m$
   Compute $\gamma_{jm} = \frac{1}{2} \frac{\sum_{x_i \in R_{jm}} r_{im}}{\sum_{x_i \in R_{jm}} |r_{im}|(1-|r_{im}|)}$,
   Update $f_m(x) = f_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$
3. Output $\hat{f}(x) = f_M(x)$

Thus MART fits small trees (built by CART) to the current stage's generalized residuals (derivative of the loss function) at each iteration. The resulting model is a weighted combination of a large number of simple trees. We emphasize that this is but one choice for the classification machine and many other choices would be suitable. For example, we have also had good success with variations of SVMs.

2.3 Variable selection/relevance/importance

As a result of logistic modelling above we get relatively complex decision regions and an important question rises: out of possibly hundreds variables which are important and to what degree with respect to defining the in-control joint density region.

For tree based models there is relatively simple way to estimate relative variable importance. For a single decision tree Breiman et al. [12] proposed

$$M(x_m, T) = \sum_{t \in T} \Delta I(x_m, t) \tag{8}$$

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

where $\Delta I(x_m, t)$ is the decrease in impurity due to an actual (or potential) split on variable x_m at a node t of the optimally pruned tree T. The sum in (2) is taken over all internal tree nodes where x_m was a primary splitter or a surrogate variable. For any primary splitter CART keeps a number of surrogate variables that mimic primary split [12].

For additive trees such as MART this importance measure is easily generalized. Simply average over the trees to

$$M(x_i) = \frac{1}{M} \sum_{m=0}^{M} M(x_i, T_m) \tag{9}$$

Due to the stabilizing effect of averaging, this measure turns out to be more reliable than it its counterpart for a single tree. Also because of its regularization strategy the masking of important variables by others with which they are highly correlated is much less of a problem. Therefore, in (2) sum is evaluated over internal nodes where the variable of interest is the primary splitter.

Note that using the same framework one could easily extract independent components of multivariate data set, and either control them separately (using simple univariate SPC) or just ignore them (if there is no known impact on quality indicators). To do that the reference distribution would be an independent version of the base data set and the sample is generated by randomly permuting values for each variable in the base data set as mentioned previously. After a binary classification model is built, variables with low relative importance signal their independence from the joint density.

2.4 Diagnosis

Given a model to assign a new observation to in/out-of-control state, and a set of importance scores for controllability variables, it would be very useful to obtain insight on the nature of the fault.

Let $\mathbf{O}$ be an outlier point, and consider the value of

$$M_i(\mathbf{O}) = \max_{x_i} P\{\mathbf{X}\ in\ control \mid x_j = o_j, \forall j \neq i\} \tag{10}$$

$M_i(\mathbf{O})$ is the maximum probability for a point $\mathbf{O}$ to be in control along dimension x_i where the rest of coordinates o_j are fixed. This is easy to calculate given the model (via a simple location search). If there is an i such that $M_i(\mathbf{O})$ is large enough (say > 0.5) it would imply that O is x_i-outlier (out-of-control in dimension x_i). Small values of $M_i(\mathbf{O})$ for all coordinates would indicate that O is a global outlier, and could not be brought back in control by adjusting any single variable. The distance along dimension x_i between the i^{th} coordinate o_i and point where $maximum$ $M_i(\mathbf{O})$ is achieved provides a measure of "outlierness" in the i^{th} dimension.

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

3 Illustrative example

To demonstrate the proposed method we use a challenging data set for a non-parametric approach. That is, a data set in which traditional assumptions are valid. We need to capture the smooth, closed elliptical boundary of a two-dimensional normal distribution with a nontrivial covariance. This model is our pattern of standard (in-control) operations.

Figure 1 shows example the in-control data generated from the two-dimensional normal $\mathbf{X} = \mathbf{C} * \mathbf{Z}$ with covariance $Cov(\mathbf{X}) = \mathbf{C} * \mathbf{C}' = \begin{pmatrix} 2 & -6 \\ -6 & 50 \end{pmatrix}$ with $E(\mathbf{X}) = \mathbf{0}$.

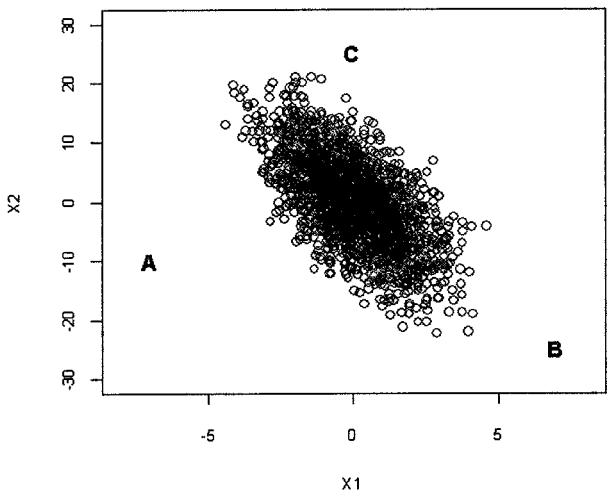

Figure 1: Two-variate normal in-control distribution with 3 outliers.

The MART algorithm was used to build a model for $p(\mathbf{x})$, the probability to be in-control given an observation vector $\mathbf{x}$. We used base trees of size 10, regularization parameter $\gamma = 0.1$, and the maximum number of iterations $M = 400$. Generated data was divided into training (70%) and testing (30%) data sets, and the sum of trees with minimum test error was used as a final model (approximately 150 trees). Figure 2 shows contour plots of the learned in-control probabilities with very impressive generalization accuracy for a generic, non-parametric classifier of 82%.

Figure 3 illustrates the points made in evaluation of "outlierness" section. It shows dependencies

$$f(x_i) = P\{\mathbf{X}\ in\ control \mid x_j = o_j, \forall j \neq i\}$$

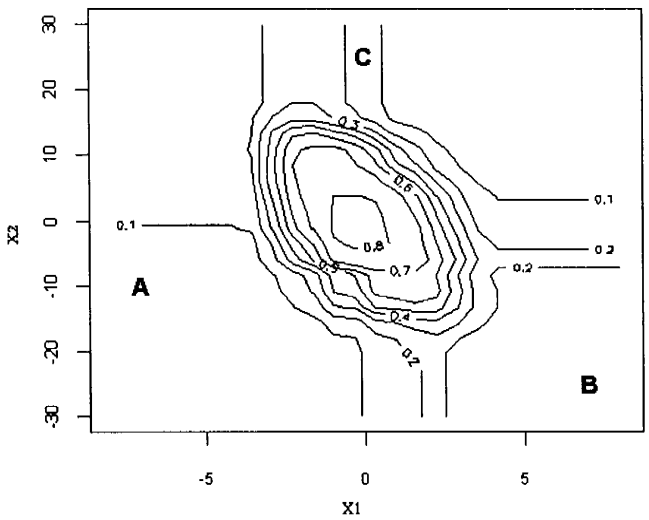

Figure 2: Learned contour plots of in-control probability.

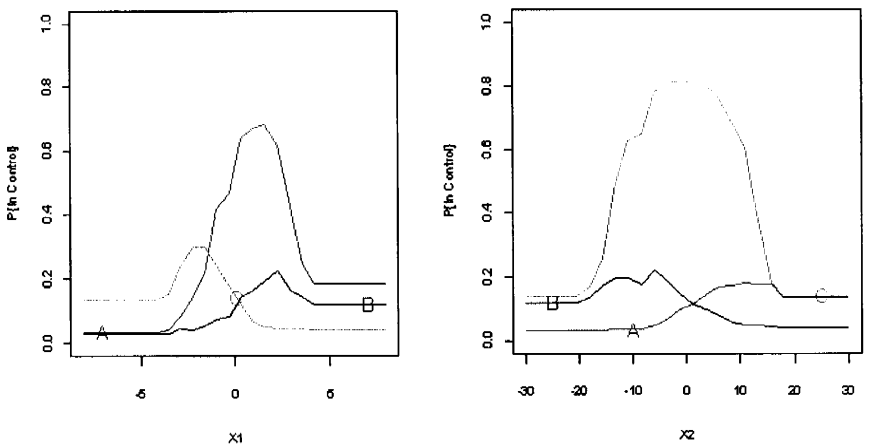

Figure 3: Conditional in-control probability for three outliers along $X_{1/2}$ dimension given the second coordinate is fixed.

for three different outliers: $\mathbf{A}(x_1 - outlier)$, $\mathbf{C}(x_1 - outlier)$, $\mathbf{B}(global\ x_{1/2} - outlier)$.

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

4 Conclusions and future work

Our nonparametric approach to learning patterns can be used for a wide range of problems. A solution to the difficult challenge to incorporate heterogeneous data types has been proposed. As shown in the figures, the nonparametric boundary captures the form of the ellipse in the test case of normally distributed data. We have demonstrated it is feasible to apply the well-developed theory for classification to this problem. Consequently, several solution methods are available. For example, in related work we used support-vector machines for similar purposes, with similar success. Our results are preliminary but promising.

Important questions remain. The control of false alarms and the stability of the estimated boundaries as a function of training sample size must be studied. The effect of dimensionality and the extension to higher dimensions is critical.

In other future work we will consider supervised pattern learning in the presence of a global quality characteristic (most often a binary characteristics like "pass"/"fail") that could be fully/partially explained or completely independent from the variables under consideration. In practice a subset of the "pass" portion of the data is considered to be a standard, and a reference set is formed against it. Then the "fail" subset of the data forms the third class. A higher cost of misclassification of "fail" class to the standard class could be used in the model training.

References

[1] Hotelling, H. *Multivariate quality control-illustrated by the air testing of sample bombsights, in Techniques of Statistical Analysis*, C. Eisenhart, M. W. Hastay, and W. A. Wallis, eds., New York: McGraw-Hill, pp. 111-184, 1947.

[2] Jackson, J. E., *Principal Components and Factor Analysis: Part I Principal Components*, Journal of Quality Technology 12, pp. 201-213, 1980.

[3] Runger, G. C., Alt, F. B., and Montgomery, D. C., *Contributors to a Multivariate Control Chart Signal*, Communications in Statistics - Theory and Methods, 25, pp. 2203-2213, 1996.

[4] Friedman, J. H., Hastie, T. and Tibsharani, R. *Additive Logistic Regression: a statistical view of boosting*, Annals of Statistics, 28, pp. 337-307, Stanford University, 2000.

[5] Vapnik, V. N., *Statistical Learning Theory*, Wiley, New York, 1998.

[6] Friedman, J. H., *Greedy Function Approximation: a Gradient Boosting Machine*, Technical report, Dept. of Statistics, Stanford University, 1999.

[7] Friedman, J. H., *Stochastic gradient boosting*, Technical report, Dept. of Statistics, Stanford University, 1999.

[8] Breiman, L., *Random Forests,* Machine Learning, Vol. 45, 1, 2001.

[9] Cucker, F., and Smale, S., *On the mathematical foundations of learning*, Bulletin of AMS, 39:149, 2001.

[10] Poggio T., Smale, S., *The Mathematics of Learning: Dealing with Data*, Notices of the AMS, May 2003.

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

[11] Freund, Y. and Schapire, R., *A decision-theoretic generalization of on-line learning and an application to boosting,* In Proc. of the Second European Conference on Computational Learning Theory, LNCS, March 1995.

[12] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., *Classification and Regression Trees,* Belmont, CA: Wadsworth, 1984.