TARGET-AWARE ANOMALY DETECTION AND DIAGNOSIS

Alexander Borisov[1], George Runger[2] and Eugene Tuv[1]
[1]*Intel, Chandler, AZ, U.S.A.*
[2]*Arizona State University, Tempe, AZ, U.S.A.*

Keywords: Outliers, Process control, Contribution plots, Partial least squares.

Abstract: Anomaly detection in data streams requires a signal of an unusual event, and an actionable response requires diagnostics. Furthermore, monitoring for process control is often concerned with one or more target (controlled) attributes. Consequently, it is necessary to separate anomalies (and their contributing attributes) that could influence the controlled target strongly, and this becomes more important with the increased number of monitored attributes in modern processes. This task leads to a difficult problem not addressed directly by the machine learning/process control community. We introduce the target-aware anomaly detection problem and present a solution for process control in modern systems (with nonlinear dependencies, high dimensional noisy data, missing data, and so on). The main objective is to identify and rank outliers and also diagnose their contributing attributes with respect to the possible effect on the response. The method is different from traditional linear and/or univariate approaches, as it can deal with local data structure in the neighborhood of an outlier, and can handle complex interactions via the use of an appropriate learner. In addition, the method can be computed quickly and does not require time consuming matrix operations. Comparisons are made to traditional contribution plots computed from partial least squares.

1 INTRODUCTION

The importance of anomaly detection has grown from manufacturing to include systems such as environmental, security, health, supply chains, transportation, etc. Data are characterized by a set of input attributes (or variables), and one (or several) controlled (target) attribute(s). A typical example of an input set consists of measurements generated over different stages of a production process from numerous sensors (from hundreds to thousands). Along with the input measurements target attributes may result from final product tests or other system performance measures. The usual goal of the process control is to keep the values of the target attributes within a given range. Furthermore, if we observe or predict an excursion in a target, it is important to determine what inputs contribute to the excursion.

In applications (such as process control) where the objective is to control a target attribute, and as the number of input attributes increase, it becomes important to focus anomaly detection to the effect on the target. Furthermore, although target-labeled training data is used, we do not assume that the process data is target-labeled when the outlier diagnostics are computed. For many common process control applications the decision for action must be based on the input attributes before the target is measured. Consequently, we refer to our approach as target-aware, rather than supervised.

This problem has not been addressed in the machine learning literature. Instead, commonly used methods in advanced process control applications (for both unsupervised and supervised anomaly detection) rely on linear data models (i.e., principal component analysis (PCA) and partial least squares (PLS) (Hastie et al., 2001)) and multivariate normal distribution assumptions in the input space. Such approaches (e.g., for example, (Ergon, 2004)) can be successful in finding anomalies in low-dimensional, non-noisy numeric data when the dependency between the inputs and outputs are near-linear. PLS provides an approach for target-labeled outlier detection and attempts to identify input attributes that most strongly contribute to an outlier. However, the above strong assumptions limit the applicability for data that can be very large (millions of samples and/or several hundreds to thousands of predictors), noisy, with heterogeneous type (numerical and categorical). Other anomaly detection approaches (e.g., see reviews by (Hodge and Austin,

Borisov A., Runger G. and Tuv E..
TARGET-AWARE ANOMALY DETECTION AND DIAGNOSIS.
DOI: 10.5220/0003530100140023
In *Proceedings of the 8th International Conference on Informatics in Control, Automation and Robotics* (ICINCO-2011), pages 14-23
ISBN: 978-989-8425-74-4

2004)) can not be easily extended to supervised settings.

Solutions to the problem above can be approached in several ways. One is to build a model on previously seen data that is capable of predicting failures or large changes in a target from recently arrived new data. We refer to this as predictive modeling. Such an approach based on different types of regression models was used in an off-line setting (Angelov et al., 2006). However, an important concern is that such models may generalize poorly on new data when the distribution of inputs changes. To address this concern previous work considered adaptive models were updated based on new test data that were determined to be fault-free (Lughofer and Guardioler, 2008; Filev and Tseng, 2006). An advantage of our approach is that the model need not be updated and we integrate unsupervised information.

An alternative is unsupervised anomaly detection where the objective is to find points in input space that are distant from the ”general” input distribution, thus effectively detecting anomalies in new (or existing) data (Chiang et al., 2001). However, such an anomaly in the input attributes might (but does not always) lead to a large change in the target value too. Still there is not a direct link between the two, as the unsupervised nature of anomaly detection does not take into account the predictive model that is used.

We propose a novel target-aware anomaly detection approach that combines both predictive modeling and unsupervised outlier detection. We use a predictive model learned from target-labeled training data to assign ”outlier scores” to new data points. These scores consider both the effect on the target as well as the remoteness of point in the space of inputs. (Note that this is different from predicting target probabilities from the model directly, as those probability estimates also do not generalize outside of the current input distribution.) Furthermore, we propose a measure that separates the most influential outliers, and ranks them according to influence on the target, along with a threshold used to filter the non-influential outliers and normal samples. As an important by-product for signal diagnosis, we compute the contribution of input attributes to the anomaly score. This task is related to fault isolation and several methods have been used (Efendic et al., 2003; Runger et al., 1996). We propose a ranking measure, along with a threshold, to identify the input attributes that contribute to a particular point being an outlier. Our analysis statistically quantifies risks and we use quite different models and metrics than previous work.

Regarding terminology, we note that supervised and unsupervised anomaly detection can refer to whether samples are labeled as *normal* or *anomalous* and not the target attribute labels (Chandola et al., 2009). Consequently, we refer to our method as target-aware anomaly detection. We performed experiments with simulated data sets that closely mimic the behavior of real systems and show that our method outperforms classical PLS contribution plot analysis, both in terms of detecting outliers, calculating relevant contributions, and in terms of improving predictive accuracy on new data when potential outliers are detected automatically even before the target is known. The algorithm was successfully applied on real data too. Here we focus on the case of a numerical target (regression case) with numerical inputs because it is more relevant for statistical process control. We provide a short description of traditional PLS based anomaly detection approach in Section 2. Section 3 presents our new method. Section 4 offers illustrative examples and compares to the traditional PLS contribution plots, and Section 5 provides conclusions.

2 TRADITIONAL LINEAR METHODS FOR SUPERVISED ANOMALY DETECTION

The supervised outlier detection and contribution problem has been addressed in the statistical literature through linear models with PLS regression (Wold et al., 2001). The PLS approach is the only method that has been been applied for outlier/contribution analysis in the supervised case. PLS iteratively computes components (directions) with maximum correlation to the target, performing Gram-Schmidt orthogonalization of the input space and the target with respect to the new identified component. The process continues until a predefined number of components is generated. In this way it yields a number of mutually orthogonal directions that have maximum correlation to the target. Projections of the input space along these directions yield loadings/projection/scores (similar to PCA). PLS results in the regression coefficients for target y with respect to these directions.

More formally, a PLS model has the form $X = TP^t$, $y = TQ^t$, where P, Q are X and y loadings, respectively, and t denotes matrix transpose. Here T is the score matrix with columns corresponding to directions. So in addition to finding a new basis in X space it provides a regression model for y. Similar to PCA, a PLS model is sensitive to the scaling of attributes, and therefore requires an appropriate scale

to be selected. In most applications the attributes are standardized (zero mean and unit standard deviation) and for simplicity we make this assumption in this paper. The PLS algorithm description can be found, for example, in (Hastie et al., 2001). The number of directions is usually selected using an explained variance threshold or cross-validation.

Two common statistics are used to monitor for anomalies (similar to the PCA case). First is squared prediction error (SPE). That is, the distance from an inspected sample to the model plane defined by selected latent variables. Second is the distance from zero (the center of the scaled distribution) in score space. To account for different score scales, and their different influence on the target, scores are multiplied by the corresponding y-loadings and divided by the standard deviation. More insight on these two measures follows next.

Suppose we computed and selected K directions $\{T_1, T_2, \ldots, T_K\}$ from a PLS model given by $X = TP^t,\ y = TQ^t$, where $T = T(N \times K),\ P = P(N \times K),\ Q = Q(1 \times K)$ matrices. Let the kth column of T and Q be denoted as T_k and Q_k, respectively. Here Q denotes the coefficient estimates from the linear regression of y on T and the (scalar) Q_k is a measure of the weight of T_k in the regression. A modified Hotelling's T^2 statistic (Hotelling, 1947) with respect to selected components is computed as follows:

$$T^2 = \sum_{k=1}^{K} \left[\frac{T_k \cdot Q_k}{\lambda_k} \right]^2, \qquad (1)$$

where λ_k is the standard deviation of the elements of T_k. Here T^2 is similar to the (Mahalanobis) distance of data sample $\boldsymbol{x}_0$ from the centroid $\bar{\boldsymbol{x}}$ after a projection to the subspace defined by the first K latent variables. An anomaly is signaled if this distance is too large. However, multiplication by the y-loadings Q allows one to take into account the influence of components on the target.

Because T^2 statistic is not sensitive to anomalies that are far from the subspace of the latent variables, a second statistic is used that is sensitive to the distance from this subspace. Suppose $R = R(N \times K)$ is the projection matrix from the original input space to scores, i.e $T = XR, P^tR = I$. The squared prediction error (SPE) is

$$SPE_0 = (\boldsymbol{x}_0 - \hat{\boldsymbol{x}_0})^t (\boldsymbol{x}_0 - \hat{\boldsymbol{x}_0}) \qquad (2)$$

where $\hat{\boldsymbol{x}_0} = \boldsymbol{x}_0 RP^t$ is the projection of $\boldsymbol{x}_0$ to the subspace spanned by the first K PLS components in the input space.

We define the (supervised) PLS contribution score of attribute j to data sample $\boldsymbol{x}_0$ to T^2, in a manner similar to (Ergon, 2004) and (Miller et al., 1998), but scores are multiplied with their corresponding y-loadings. That is,

$$C_j(T^2, x_0) = x_{0j} \cdot \sqrt{\sum_{k=1}^{K} \left(\frac{T_k Q_k \cdot R_{jk}}{\lambda_k} \right)^2} \qquad (3)$$

and this can be interpreted as the multiplication of the scores (normalized and weighted by y-loadings) by the term $x_{0j}R_{jk}$ that gives influence of attribute j on the score $T_k = XR_k$.

Similarly, the PLS contribution score of attribute j for $\boldsymbol{x}_0$ to SPE is calculated from the j-th term of (2) in the same way as for PCA model. That is,

$$C_j(SPE, x_0) = (x_{0j} - \hat{x}_{0j})^2 \qquad (4)$$

3 TARGET-AWARE ANOMALY AND CONTRIBUTOR SCORING ALGORITHM

As stated earlier, the notion of "target-aware outlier/anomaly detection" is closely related to the effect on a target. More precisely, call an outlier in X space an *influential outlier* if it has a large influence on the target. Call all other outliers (with minor or no influence on the target) *non-influential outliers*. For example, an outlier in an irrelevant attribute will have no influence. Our goal here is to detect influential outliers and determine the attributes that contribute to them. We assume training data with target-labels for a predictive model, but to meet the conditions in process control applications we assume that target values are not available at the time anomaly detection is evaluated.

Due to the complexity and size of real data sets, and to address nonlinear models, interactions, missing data, different units for attributes, and other properties of practical problems, we develop new scores and diagnostics from tree based models. An ensemble of gradient boosting trees models (GBT) (Friedman, 2001) is used to calculate a "target-aware outlier score", that describes how well a new (or old) sample fits into the input distribution partitioned by trees in the model with adjustment for the target.

3.1 Decision Tree Formulas

We consider the practical case where both the input attributes and the target are continuous. A decision tree provides a supervised predictive model (Breiman et al., 1984) that recursively partitions samples to achieve smaller impurity of the target attribute. The samples at a node are partitioned (split) into subsets at

the two child nodes (because we only consider binary trees). Each split is defined by a binary rule on one of the input attributes X_j. For a continuous attribute partitions of the form $X_j < v$ or $X_j > v$ are considered. The attribute for which the binary rule minimizes the impurity of the target is called the primary (best) splitter. The impurity is measure defined by entropy or the Gini index for categorical targets and for a numerical target we use squared error loss

$$I(T) = \sum_{i \in T} (y_i - \bar{y}(T))^2$$

where $\bar{y}$ is the mean of the target values at the node. After a split the impurity of the two child nodes is computed as

$$I(T_L) + I(T_R) = \sum_{i \in T_L} (y_i - \bar{y}(T_L))^2 + \sum_{i \in T_R} (y_i - \bar{y}(T_R))^2$$

where $\bar{y}(T_L)$ and $\bar{y}(T_R)$ are the target means in the left and right child nodes, respectively. The change in impurity from the parent to the child nodes obtained from the primary splitter is denoted as

$$W(T) = I(T) - I(T_L) - I(T_R) \quad (5)$$

and referred to as the split weight. Larger values indicate a more important split at node T.

The process continues until a stopping rule is true. A single tree algorithm might further prune (remove) nodes, but typically ensemble models do not. Given a sample x_0 the attribute values in x_0 determine a path through the nodes until a terminal (leaf) node is reached. The majority class of the target at the leaf is often used to predict a categorical target and the mean of the target values at a leaf is often used to predict a continuous target. A GBT model is a (serial) ensemble of decision trees that is used to improve predictive performance (Friedman, 2001).

3.2 Target-aware Algorithm

The algorithm starts with target-labeled training data and a GBT ensemble generates a predictive model to relate the process (input) attributes $(X_1, X_2, \ldots, X_M)$ to the target attribute Y. In the test phase the objective is to determine if a new sample x_0 is an influential outlier, without the label Y. One may often have a collection S_0 of samples in the test data so that ranking and selection of the influential outliers, along with attribute contributions, are important. In the following, for simplicity, we describe the scoring for a single test sample x_0. We calculate the contribution of each attribute X_j to the outlier score for sample x_0 and this simply leads to the final outlier score as the sum of the contribution scores from all attributes. The contribution of X_j depends on the remoteness of the value of X_j in sample x_0 in X-space (the X-contribution), as well as the importance of X_j to predict the target (the Y-contribution). Measures for both these terms can be calculated quickly from the previously learned GBT model.

For the X contribution, consider node T of a tree in the ensemble. We define

$$d_j(T, x_0) = \frac{|x_{0j} - \bar{x_j}(T)|}{IQR(X_j, T)}$$

where $\bar{x_j}(T)$ is mean value of X_j in the node T (computed from the non-missing values), and $IQR(X_j, T)$ is the difference between the 75% and 25% quantiles of X_j in the node T. The IQR is proportional to a robust estimate of the standard deviation. Here $d_j(T, x_0)$ measures the remoteness of the value of X_j in sample x_0 within the data at node T. The IQR is used in a similar manner to a standard deviation, to (robustly) adjust for different units among the X_j's. The algorithm described below computes

$$\max[0, d_j(T, x_0) - C_1]$$

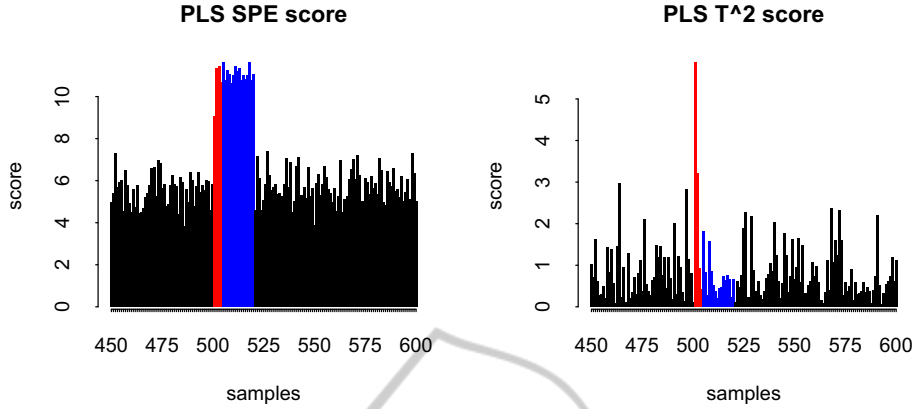
for a predefined (tuning) constant C_1. If the contribution is negative it is set to zero. The role of C_1 is to truncate small absolute differences to zero. Small differences indicate that x_{0j} is not remote at node T.

For the Y contribution, it is useful to compare the split weight in equation 5 actually obtained at a node T to a baseline. The baseline is the split weight obtained after the target values at node T are randomly permuted among the samples (denoted as $W_0(T)$) and minimum impurity is computed. That is, for the rows at node T the y measurements are randomly permuted among these rows, and then the minimum split weight is determined as in 5. The intent is that is $W_0(T)$) provides a baseline score for split weight when inputs are not related to the target. Consequently, in our outlier algorithm we consider an adjustment to the split weight as

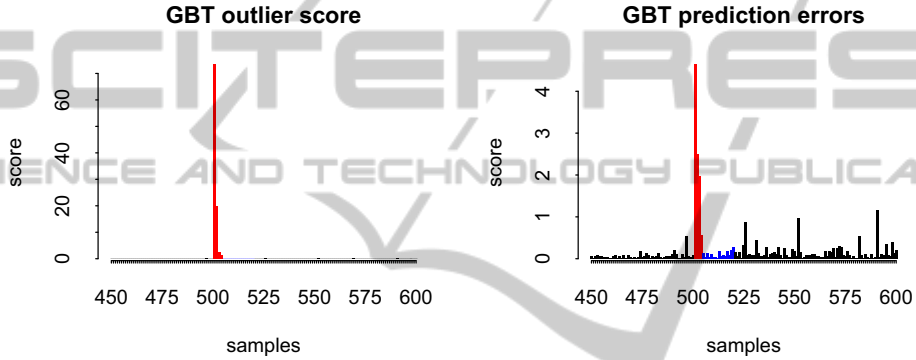
$$\max(0, \sqrt{W(T)} - C_2\sqrt{W_0(T)})$$

where C_2 is a (pre-specified) tuning constant. The role of C_2 is to truncate Y-contributions that do not exceed a baseline to zero. We apply a square root to put both X- and Y-components on the same linear scale. The importance of X_j to predict the target is simply obtained from a function of the split weight $W(T)$ (and the baseline $W_0(T)$) at the nodes where X_j is the primary splitter.

Then the contribution of the attribute in the node is computed as the product of the X- and Y-contributions. This allows us to take into account the "local" effect of attribute effects on the target in a manner similar to the PLS T^2 score. But PLS uses



(a) SPE, T^2 scores from PLS.



(b) Outlier scores from GBT and prediction errors on the samples (for reference).

Figure 1: *SPE* and T^2 score plots from PLS vs. our target-aware method for experiments with a basic linear model. Influential and non-influential outliers are shown with red and blue bars, respectively.

the Y-loadings as multipliers, i.e., "global" score multipliers. Both constants are usually fixed. We used the same values $C_1 = 3$, $C_2 = 2.5$ for all experiments. The algorithm is not too sensitive to C_1, and C_2 acts as false alarm/rejection rate threshold to detect influential outliers.

The details of the algorithm follow.

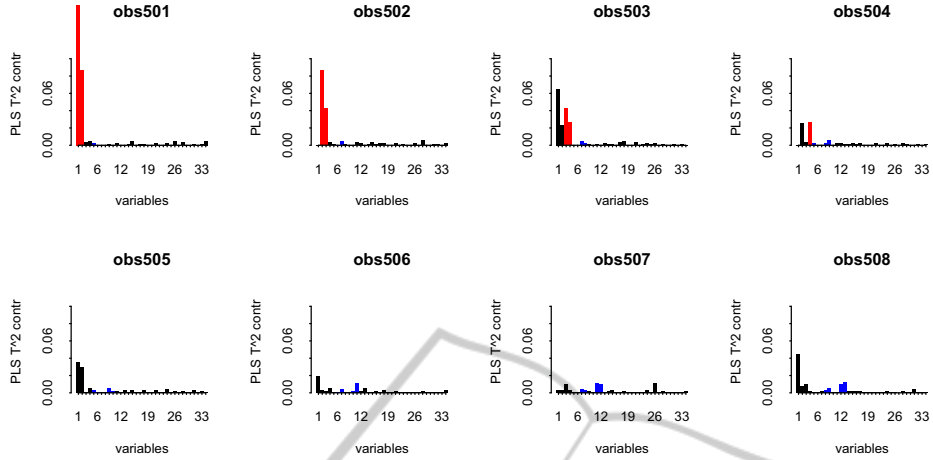
1. Build a GBT model for a given target. Initialize the contribution score matrix $C_j = 0,\ j = 1 \ldots M$.
2. Compute the contributions of an attribute to a selected sample x_0. The contribution calculation is based on the GBT ensemble. Select an attribute X_j. For each node T in each tree, where the primary splitter is X_j, the contribution $C_j(T)$ of variable X_j to sample x_0 is defined as

$$\begin{aligned} C_j(T) &= \max(0, d_j(T,x_0) - C_1) \\ &\quad \cdot \max(0, \sqrt{W(T)} - C_2\sqrt{W_0(T)}) \end{aligned}$$

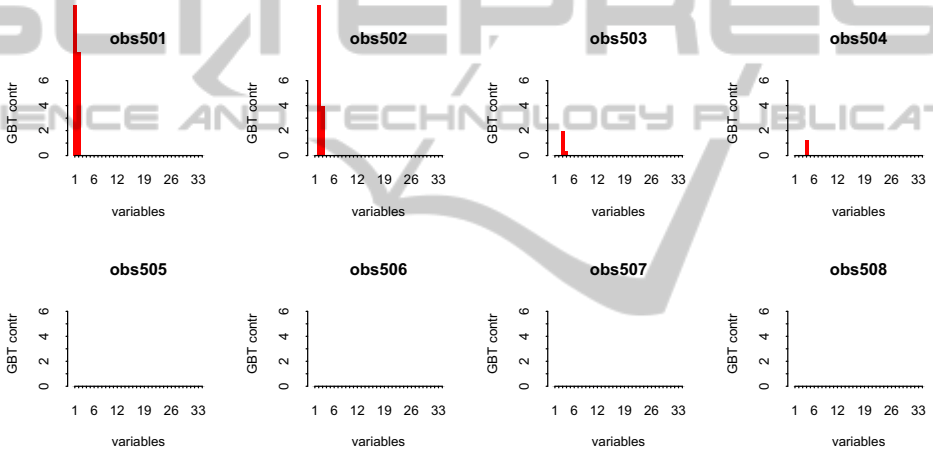
   Then the contribution score C_j of variable X_j to sample x_0 is increased by the term $C_j(T)$.
3. Compute the total outlier score for sample x_0 sums over all attributes as

$$C_0 = \sum_{j=1}^{M} C_j$$

The most time consuming operation of our method is building a GBT model, but this is only performed once (or when the model is refreshed). As each tree has time complexity $N \cdot \sqrt{N} \cdot M$ we obtain $O(K \cdot N \cdot \sqrt{N} \cdot M)$ complexity for GBT model with K trees. For very wide data sets we can use embedded feature selection for GBT (Borisov et al., 2006), and the complexity is reduced by a factor of $\sqrt{M}$. Although it is still more expensive than the PLS algorithm, we avoid a quadratic complexity in both the number of samples and the number of attributes. To score a sample x_0 as a potential outlier the steps are comparable to a prediction from a GBT model (traverse the trees in the model), along with some simple intermediate calculations for the X- and Y-contributions, and these computations are fast on modern hardware for even large data sets.



(a) Contributions from PLS T^2 scores.

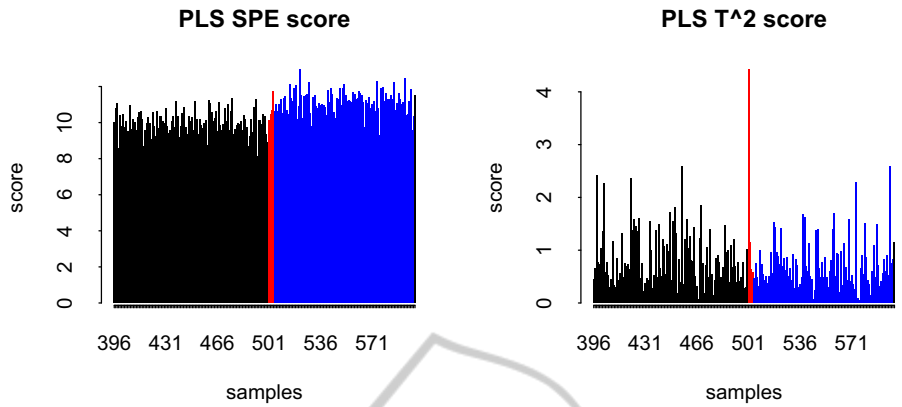


(b) Contributions from the target-aware outlier scores.

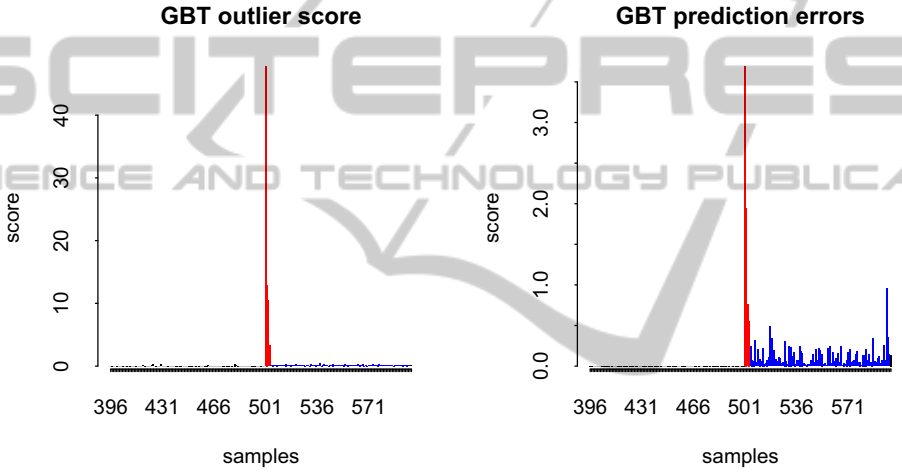
Figure 2: T^2 score contribution plot from PLS vs. contribution plot from our target-aware method for experiments with a basic linear model. Contributions for all variables are shown for the first 8 test samples. Correct relevant contributors are shown with red bars. Other contributors are shown with blue bars. For the first three outliers there are 2 relevant contributors, the fourth outlier has only 1 relevant contributor, and the other four groups have none.

4 EXPERIMENTS

The only existing competitive method for supervised outlier ranking and contribution estimation is a PLS contribution plot analysis which is widely used in statistical process control (along with more traditional PCA). However, linear methods can be ineffective for nonlinear high-dimensional data with many noisy or irrelevant attributes. We illustrate the limitations of PLS on several experimental scenarios that are very similar to actual problems that are encountered in manufacturing applications. Typical data have very few relevant inputs and/or contributors for an outlier, even when the total number of inputs can range from several hundreds to tens of thousands. Therefore, consider similar, representative scenarios in our experiments, but with the pragmatic advantage that the ground truth is known. From a linear model built with PLS we can obtain two scores: SPE (distance from the model that is computed in the same way as PCA) and T^2 (modified with Y-loadings). It is worthy to mention that the SPE score is useless for distinguishing influential/non-influential outliers or contribution analysis as it does not use Y-loadings.



(a) SPE, T^2 scores from PLS.



(b) Outlier scores from our target-aware method and prediction errors on the samples (for reference).

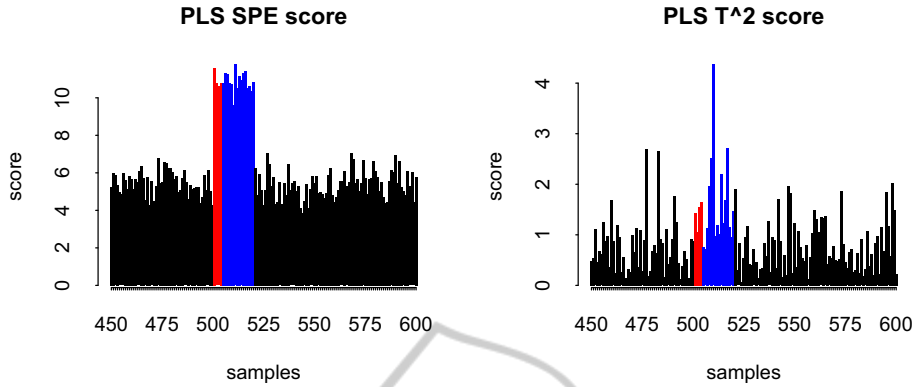
Figure 3: *SPE* and T^2 score plots from PLS vs. our target-aware method for a linear model with 100 noise variables. Samples with relevant contributors (influential outliers) and non-influential outliers are shown with red and blue bars, respectively. Each outlier has one contributor.

4.1 Linear Data with Noise

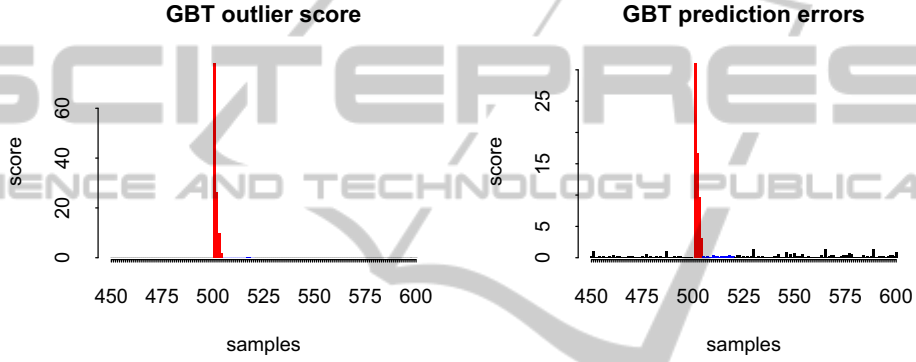
Reference data are simulated from 34 independent attributes with 500 samples (rows). Each is normally distributed with mean zero and standard deviation one. The four first attributes are relevant predictors, with the target $y = x_1 + 0.5x_2 + 0.3x_3 + 0.15x_4$. The other 30 attributes are irrelevant. We added 100 test samples generated from the same distribution, in which we create 20 anomalies. Anomaly i has values of attributes $x_i, x_{i+1}, x_{i+4}, x_{i+5}$ equal to 5. Therefore, the first three outliers have 2 relevant and 2 irrelevant contributors, the fourth outlier has 1 relevant contributor, and the other outliers have 0 relevant contributors (all 4 contributors irrelevant) to the target. The targets for these additional samples are generated from the same linear function of relevant variables. For PLS we used two latent components as suggested by cross validation based on the explained y-variance. (The two components explained 98% of the y variance.) A GBT model was built with 700 trees, shrinkage rate = 0.01, tree depth = 4, and each tree used 60% subsampling. The least squares loss function was employed.

Figure 1 shows the PLS SPE and T^2 plots for 50 (last) training samples and all testing samples for this example, compared to the GBT outlier scores plot.

Figure 1 shows that PLS cannot distinguish the influential and non-influential outliers. Actually it does not even separate the two last influential outliers, while the target-aware outlier score plot generates the correct results. All four influential outliers are



(a) SPE, T^2 scores from PLS.



(b) Outlier scores from our target-aware method and prediction error on the samples (for reference).

Figure 4: *SPE* and T^2 score plots from PLS vs. our target-aware method for nonlinear data. Samples with relevant contributors (influential outliers) and non-influential outliers shown with red and blue bars, respectively.

detected, without false alarms. Furthermore, the influential outliers are correctly ranked and their scores are higher than the scores for the non-influential outliers and the normal samples (non-outliers). Also, the figure shows that the target-aware outlier scores are very well correlated to prediction errors for the corresponding samples. Although, as mentioned previously, the prediction errors are assumed to not be available at the time of these diagnostics. The PLS T^2 plot cannot even separate outliers from non-outliers. While the SPE plot can separate in this case, below we have the slightly modified version of this example where it fails too.

Figure 2 shows plots of PLS T^2 contribution scores and target-aware outlier score contributions for the first 16 test samples.

Again PLS does not correctly rank and separate the relevant and irrelevant contributors. The target-aware outlier scores correctly detect and rank both the outliers and the contributing attributes with respect to their effect on the target.

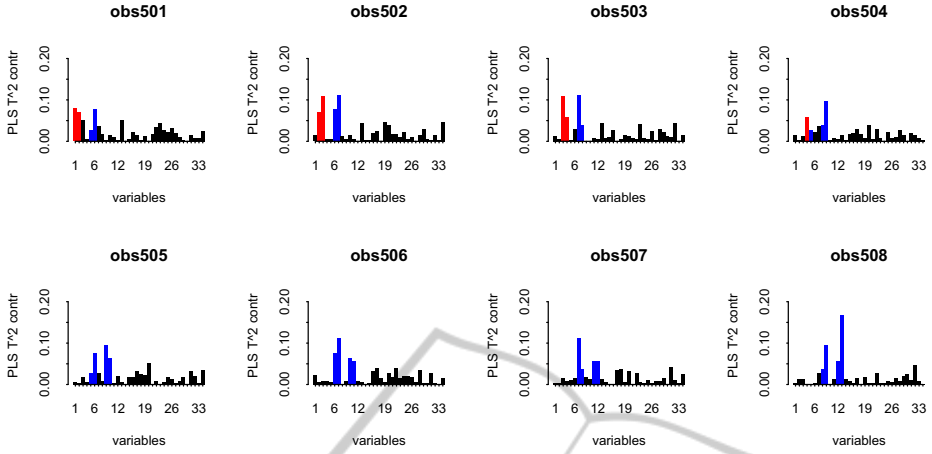
However, when the number of noise attributes increase and the outliers do not deviate much from zero, the SPE and T^2 plots may not detect them at all. Consider a simplified version of the above linear example with 100 noise attributes, and each i-th outlier is created by setting $x_i = 5$. Figure 3) shows PLS SPE and T^2 plots vs the target-aware scores plot for this example.

It can be seen that now SPE cannot readily identify outliers, although their scores are slightly above average, and T^2 shows only the single most influential outlier.

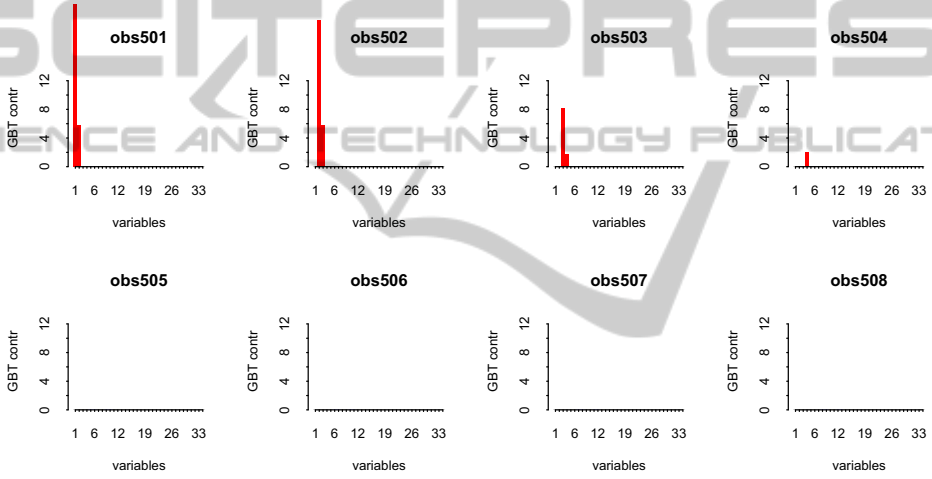
4.2 Nonlinear Data

Next we consider a similar example with the target a nonlinear function of inputs–a quadratic target function $y = x_1^2 + 0.5x_2^2 + 0.3x_3^2 + 0.15x_4^2$. All other settings were similar to the previous experiment, except we used four latent components for PLS (again as suggested by a 10-fold cross-validation model error plot).

Figure 4 shows PLS SPE and T^2 plots for 35 (last) training samples and all test samples for this example,



(a) Contribution from PLS T^2 scores.



(b) Contributions from target-aware outlier scores.

Figure 5: T^2 score contribution plot from PLS vs. contribution plot for our target-aware method for nonlinear data. Contributions for all variables are shown for the first 8 test samples. Correct relevant contributors are shown with red bars. Other contributors are shown with blue bars. For the first 3 outliers there are 2 relevant contributors, the fourth outlier has only 1 relevant contributor, and the other four groups have none.

along with our target-aware outlier scores plot.

Figure 4 indicates that PLS cannot distinguish the influential and non-influential outliers (actually it does not separate even the last two influential outliers), while the target-aware outlier scores plot generates the correct results. It shows that all four influential outliers are correctly detected and ranked, without false alarms. Also it can be seen that the target-aware outlier scores are well correlated to prediction errors for corresponding samples. The PLS T^2 plot fails to separate even outliers from non-outliers.

Figure 5 shows plots of PLS T^2 contribution scores and target-aware outlier score contributions for the first 16 test samples.

Again PLS fails to correctly rank and separate relevant and irrelevant contributors. Still, target-aware outlier scores correctly rank the outliers and the contributing attributes with respect to their effect on the target.

5 CONCLUSIONS

The target-aware anomaly detection and associated contribution problem is introduced to the machine learning community and a solution for complex data from modern systems is described. New outlier scores and contributions are developed, and thresholds (de-

rived from baselines obtained from target permutations) are used to filter non-influential outliers and normal samples. The target-aware paradigm uses target-labeled data for training, but the diagnostics are calculated before the corresponding target attribute value is available (to meet the conditions for process control applications). The ensemble model allows us to deal with complex interactions in predictor space and local data structures. Linear methods based on principal components often fail to detect outliers and/or contributors in anomaly detection, especially in the presence of noise. The linear methods are also sensitive to scaling. Furthermore, linear methods rarely work when the target is a non-linear function of the inputs. Furthermore, methods such as partial least squares often fail to rank the contributors correctly and fail to separate relevant from irrelevant contributors. When the number of irrelevant variables increases it even can fail to identify influential outliers on relevant predictors. The proposed method works equally well for linear and non-linear cases in terms of diagnostics. Our method can correctly rank outliers with respect to their effect on the target, rank attributes that contribute to an outlier score, and filter non-influential and normal sample. It is insensitive to noise and ranks outliers and contributors for a data sample using a fast, robust, nonparametric technique.

ACKNOWLEDGEMENTS

This research was partially supported by ONR grant N00014-09-1-0656. We wish to thank anonymous referees for comments that improved this work.

REFERENCES

Angelov, P., Giglio, V., Guardiola, C., Lughofer, E., and Luján, J. (2006). An approach to model-based fault detection in industrial measurement systems with application to engine test benches. *Measurement Science and Technology*, 17:1809.

Borisov, A., Eruhimov, V., and Tuv, E. (2006). Tree-based ensembles with dynamic soft feature selection. In Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L., editors, *Feature Extraction Foundations and Applications: Studies in Fuzziness and Soft Computing*. Springer.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, MA.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15.

Chiang, L., Russell, E., and Braatz, R. (2001). *Fault detection and diagnosis in industrial systems*. Springer Verlag.

Efendic, H., Schrempf, A., and Del Re, L. (2003). Data based fault isolation in complex measurement systems using models on demand. In *Proceedings of the 5th IFAC-Safeprocess 2003, IFAC*, pages 1149–1154. ACM.

Ergon, R. (2004). Informative pls score-loading plots for process understanding and monitoring. *Journal of Process Control*, 14(6):889–897.

Filev, D. and Tseng, F. (2006). Novelty detection based machine health prognostics. In *Evolving Fuzzy Systems, 2006 International Symposium on*, pages 193–199. IEEE.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Hodge, V. J. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:85–126.

Hotelling, H. (1947). Multivariate quality control-illustrated by the air testing of sample bombsights. In Eisenhart, C., Hastay, M., and Wallis, W., editors, *Techniques of Statistical Analysis*, pages 111–184. McGraw-Hill, New York.

Lughofer, E. and Guardioler, C. (2008). On-line fault detection with data-driven evolving fuzzy models. *Control and Intelligent Systems*, 36(4):307–317.

Miller, P., Swanson, R., and Heckler, C. (1998). Contribution plots: A missing link in multivariate quality control. *Applied Mathematics and Computer Science*, 8(4):775–792.

Runger, G., Alt, F., and Montgomery, D. (1996). Contributors to a Multivariate Statistical Process Control Chart Signal. *Communications in Statistics–Theory and Methods*, 25(10):2203–2213.

Wold, S., Sjostrom, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130.