2018 IEEE 14th International Conference on Automation Science and Engineering (CASE)
Munich, Germany, August 20-24, 2018

Outlier Detection for Hydropower Generation Plant

Imtiaz Ahmed[1], Aldo Dagnino[2], Alessandro Bongiovi[3] and Yu Ding[4]

Abstract— **A hydropower generation plant is a complex system and composed of numerous physical components. To monitor the health of different components it is necessary to detect anomalous behavior in time. Establishing a performance guideline along with identification of the critical variables causing anomalous behavior can help the maintenance personnel to detect any potential shift in the process timely. To establish any guideline for future control, at first a mechanism is needed to differentiate anomalous observations from the normal ones. In our work we have employed three different approaches to detect the anomalous observations and compared their performances using a historical data set received from a hydropower plant. The outliers detected are verified by the domain experts. Making use of a decision tree and feature selection process, we have identified some critical variables which are potentially linked to the presence of the outliers. We further developed a one-class classifier using the outlier cleaned dataset, which defines the normal working condition, and therefore, violation of the normal conditions could identify anomalous observations in future operations.**

I. INTRODUCTION

A hydropower generation plant can be divided into many functional areas like generators, turbines, bearings etc, each of which areas can in turn be subdivided into components. Data are generated in real time from hundreds of sensors across these functional areas and instrumented equipments. Anomaly can come from various sources and can cause different range of problems. For instance, an anomaly can be overheating of bearing oil and metal components, vibrations from bearings, or low generation of active or reactive power. It is vital to identify anomaly as soon as it appears. But doing so becomes extremely challenging as data generated is of high dimensionality (i.e., too many variables).

In the literature, the anomaly detection problem is known as the *unsupervised* learning problems, because one does not have a training dataset with observations labeled as normal observations or outliers. Consequently, a supervised learning cannot be used to learn a rule to classify future observations. Intuitively speaking, outliers or anomalies are points or clusters of points which lie away from neighboring points and clusters and they seem to be inconsistent with other observations. A perfect definition of an outlier, however, does not exist. All of the outlier detection methods developed thus far are based on some assumptions, and no single unsupervised outlier detection method can perfectly classify all different types of outliers in a dataset [1].

Existing outlier detection methods can be grouped into four major schools of thoughts depending on their criteria of identifying outliers. According to the time line of their inception to the body of knowledge, the four schools are: *distance and density based methods*, *subspace based methods*, *angle based methods* and *ensemble based methods*. Each of these domains has their respective strengths and weaknesses. In the distance based methods, a point is considered an outlier if it lies further away from most of the points [2]. Instead of considering distances from all the points it would be more logical to consider the deviation from neighborhood points and lead thus to methods based on the concept of k-nearest neighbor (k-NN) [3]–[5]. One of the major downside of these distance based methods is: if the dataset has multiple clusters of varying density then they would not be able to separate local outliers (i.e., outliers only with respect to a single cluster) and normal data points successfully. Distance-based methods tend to work effectively when the dataset has clusters of similar density or no cluster at all.

In the density based methods [6]–[9], a point is considered an outlier if the density around it is considerably lower than the density around its neighbors. So these methods can handle data with clustering tendency and can identify local outliers. Both of distance and density based methods need the pairwise distances to be calculated. When the data dimension is too high, the proportional difference between the farthest point distance and the closest point distance vanishes and distances between any pair of data records become much less differentiable [1]. Consequently, the distance and density based methods does not work effectively in high dimensional data spaces.

In data spaces of high dimension, we have to consider relevant subspaces rather than considering the entire feature set. For a particular observation, relevance means the subspace in which it is different than other observations. As such, the search for outliers must be accompanied by the search for relevant subspaces. The disadvantages of the subspace based method includes the lack of an appropriate way to compare outliers identified in different subspaces, and the large number of subspaces that need to be explored.

Angle based methods are similar to the distance based methods but they were introduced with the consideration that angles are a more stable measure in high dimensions compared to distances [10]. One major limitation of this method is the high computational time it requires to calculate the angles.

Ensemble based techniques were introduced more recently, motivated in part by the frustration that no outlier detection

[1]Imtiaz Ahmed with the Department of Industrial & Systems Engineering, Texas A&M University, College Station, TX imtiazavi@tamu.edu

[2]Aldo Dagnino with ABB Corporate Research, Raleigh, NC aldo.dagnino@us.abb.com

[3]Alessandro Bongiovi with ABB, Genoa, Italy alessandro.bongiovi@it.abb.com

[4]Yu Ding with the Department of Industrial & Systems Engineering, Texas A&M University, College Station, TX yuding@tamu.edu

978-1-5386-3593-3/18/$31.00 ©2018 IEEE

techniques have been able to identify different types of outliers, while in the other part by its success in supervised learning such as bagging or boosting [11]. Researchers feel the need to combine non-compatible techniques of different types to improve the outlier detection accuracy. To ensemble, one can either use different techniques one after another on the dataset in randomly selected subspaces or one can use one suitable technique on the dataset in randomly selected subspaces for a number of iterations and then combine the result over different techniques/iterations for each observation. Using ensemble based techniques, how to combine scores from different outlier methods is still an issue elusive of the data mining community [1].

In this paper we chose three outlier detection methods based on distinct schools of thoughts, namely, *Local Outlier Factor (LOF)* as a density based method, *Feature Bagging for Outlier Detection (FBOD)* as an ensemble method, and *Subspace Outlier Degree(SOD)* as a subspace method. We employed these methods to identify the outliers in a dataset received from a hydropower generation plant. The comparative performances of these methods are analyzed and some commonality among the results are found. We discuss our finding concerning which variables have the most contribution to the selected outliers and for what range of values. We have also trained a one-class support vector machine (SVM) classifier based on the outlier-removed hydropower plant dataset. The one-class SVM defines the boundary for normalcy and can thus be used to check future observations for their likelihood of being an outlier.

The rest of the paper unfolds as follows: Section II analyzes the dataset received, how it was cleaned, and summarizes the research question at the end. Section III describes the outlier detection methods that we selected to apply on our dataset. Section IV presents the results from applying the selected methods to the hydropower dataset. Analysis of the results follows in Section V. Finally, we conclude the paper in Section VI.

II. Physical System and Dataset

The data was originated from raw data stored in the Distributed Control System of a power generation plant. It was received in time-stamped format (several months worth of data) and divided into different functional areas (turbines, generators, bearings etc.). The data was collected at 10 minute interval in each of the day. But it was not always continuous and some days from each of these months were missing. At first, we combined the data in one file across all functional areas. There were 9,508 observations (rows in a data table) and 222 variables (columns in the data table) with no missing values. Variables are mainly temperatures, vibrations, pressure, active power etc.

The first major step consisted of conducting basic preprocessing and statistical analysis in order to clean the data before applying any anomaly detection techniques. Two months had some duplicate observations (most probably due to the error in measuring sensors) and those were removed from the analysis.

A more sophisticated anomaly is found based on the monthly density curves. Month wise density curves of two vibration variables are shown in Fig. 1, in which one can observe that the November density curve has its peak far different from other six months, raising the red flag for the November data. We become more confident on our finding after we have done a cluster analysis on the dataset. It was then found that the observations from November form a clearly separate cluster with less than five percent of observations of the data set (Table I). In Fig. 2 we plotted one of those variables against active power, in which we can easily notice that the red cluster (composed of only November data) is different from the other clusters. Later we were informed by the data owner who double checked their data collection process and found that there were some mistakes in accumulating the November data. We were then provided a slightly reduced set of corrected data for the month of November. After all the preprocessing, the number of total observations (rows in a data table) is now reduced to 9,219 from 9508.

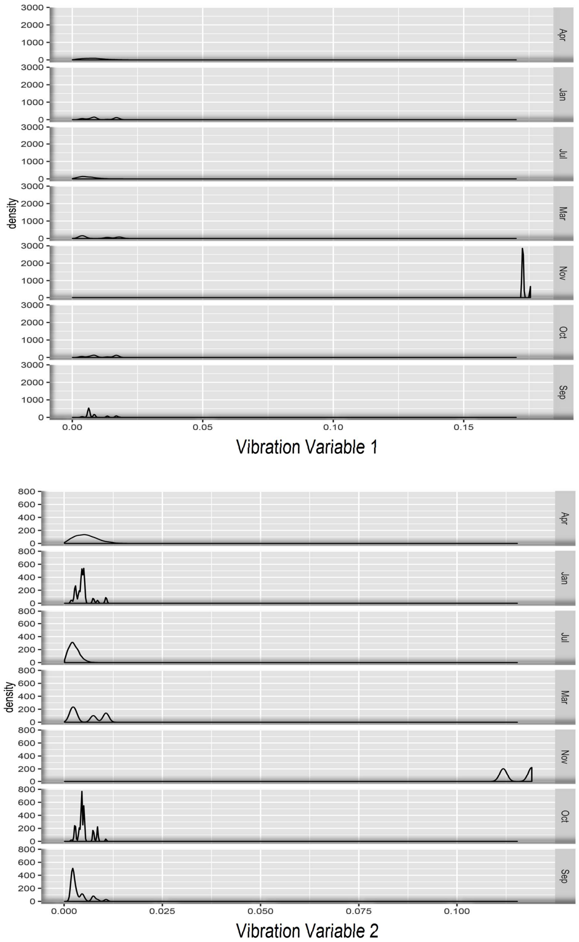

Fig. 1. Month wise Density curve for two selected variables (vibrations)

After the initial data pre-processing, we are ready to determine anomalies that are harder to detect (than the ones that are detectable via a simple clustering action). As we have said earlier, the data records are of rather high

TABLE I
CLUSTER ANALYSIS RESULT

Cluster	Apr	Jan	July	Mar	Nov	Oct	Sep
1	0	0	0	0	552	0	0
2	0	0	1565	0	0	0	0
3	662	1535	8	23	0	1527	746
4	45	413	27	1747	0	555	102

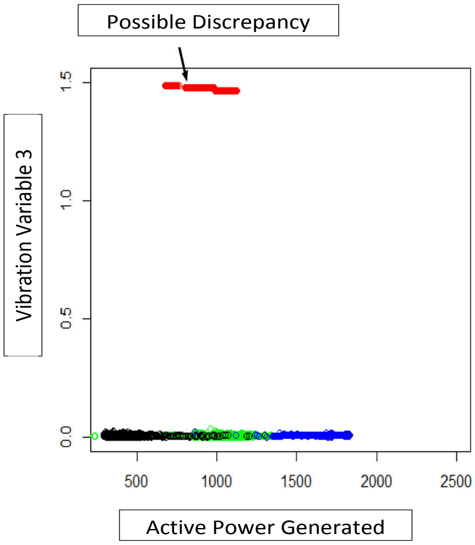

Fig. 2. Red cluster corresponding to November data

dimensions (i.e., 222). It is extremely difficult, if not altogether impossible, to analyze the data graphically for visually spotting anomalies. The data also form clusters of various densities, indicating the possibility of existence of local anomalies. To tackle the practical problems at hand, we need to utilize anomaly detection techniques that can handle this type of *unsupervised*, *high dimensional* data with *clustering tendency*. After identifying those points we also want to pinpoint the reasons and conditions which triggered these anomalous observations for the sake of assisting engineering decision making, and develop a classifier to use in the future for flagging any observations that might be an outlier.

III. ALGORITHMS USED FOR IDENTIFYING THE OUTLIERS

We choose three different outlier detection approaches and apply them to the hydropower dataset. The reason that we choose multiple methods from different schools of thoughts is because that not any one of the existing methods can address the challenging unsupervised outlier detection problem effectively. We briefly discuss the working principle of the chosen methods below.

A. Local Outlier Factor

Local outlier factor (LOF) is an algorithm for outlier detection which was proposed in [6]. They introduced a new idea called *local density*, which is evaluated by the distance of an observation from its nearest neighbors. Points which has a lower density than its neighbors (i.e., higher LOF score) will be counted as outliers. A brief outline of the algorithm is provided below

- Compute reachability distance (reach-dist) for each data point q with respect to p using (1).

$$\textit{reach-dist}(q,p) = max\{\textit{k-distance}(p), dist(q,p)\}, \quad (1)$$

where $dist(q,p)$ is the Euclidean distance between q and p and k-distance(p) is the distance from p to the k-th nearest neighbor of p

- Compute local reachability density (lrd) of q as the inverse of the average reachability distance based on the k-nearest neighbors of q as in (2).

$$lrd(q) = \frac{Cardinality\{MinPts(q)\}}{\sum_{p\in\ MinPts(q)}[\textit{reach-dist}(q,p)]}, \quad (2)$$

where MinPts(q) denotes the set of points in the k-nearest neighbors of q.

- Compute LOF(q) as the ratio of the average local reachability density of q's k-nearest neighbors and that of q, as in (3).

$$LOF(q) = \frac{1}{Cardinality\{MinPts(q)\}} \sum_{p\in\ MinPts(q)} \frac{lrd(p)}{lrd(q)}. \quad (3)$$

The dataset we are dealing with forms clusters of different density which is the ideal scenario of applying the LOF algorithm. Several variants of the original LOF method were introduced recently [7]–[9]. Though the original LOF algorithm still works better than these methods in the majority of the cases as experimented in [12]. Therefore we select LOF to apply on our dataset. One major disadvantage of LOF is still its inability to perform satisfactorily in high dimensions, especially when the outliers are different from other points only in one or very few of the dimensions.

B. Subspace Outlying Degree

To deal with high dimensional data problem, we may need to consider a subset of the original features, an action commonly known as dimension reduction. The potential benefit of looking into a subspace is that data points distributed uniformly in the full dimensional space could deviate significantly from others when examined in subspaces. This is to say, the outlierness is amplified in a properly chosen subspace. The danger of using the subspace approach is that if not chosen properly, the difference between a potential outlier and normal points may disappear altogether in another subsapce.

A good number of subspace methods are recently introduced [13]–[15]. We choose the *Subspace Outlying Degree (SOD)* [14] to be applied to hydropower dataset, as the SOD method seems not to rely on certain assumptions (such as monotonicity), arguably restrictive yet commonly used in other subspace approaches [13]. Steps of the SOD algorithm are outlined below:

- Compute the total variance (VAR) of the set of the reference points as in (4):

$$VAR = \frac{\sum_{p \in S} \left[dist(p, \mu)^2 \right]}{Cardinality(S)}, \tag{4}$$

where μ is the average position of the points in the reference set.
- Compute the variance for each variable according to (5), where μ_i is similarly defined:

$$var_i = \frac{\sum_{p \in S} \left[dist(p_i, \mu_i)^2 \right]}{Cardinality(S)}. \tag{5}$$

- Create a subspace vector based on the following criteria in (6), where d is the dimension of the original data space and α is a constant, suggested to be 0.8 in [14].

$$v_i = \begin{cases} 1, & \text{if } var_i < \alpha \cdot \frac{VAR}{d}, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

- When v_i is one, the corresponding variable is selected to construct the subspace; otherwise, the corresponding variable is skipped over.
- Finally, equation (7) calculates the SOD score as the weighted distance between an observation q and the subspace hyperplane created by the reference points of q, denoted by $\mathcal{H}(S)$. The weight, denoted by $\|v\|$, is the number of dimensions used to construct the subspace. If q deviates a lot from the reference set (shown as a higher SOD score), it is more likely to be an outlier. Reference set members are selected from the set of nearest neighbors which shares the most neighborhood similarity with q.

$$SOD(q) = \frac{dist(q, \mathcal{H}(S))}{||v||}. \tag{7}$$

C. Feature Bagging for Outlier Detection

The subspace approaches handle the high dimensionality but one of the key limitations is that if the subspace is wrongly selected, the resulting subspace could make the outlier detection harder or even not possible anymore. One approach to avoid being trapped in the wrong subspace is an ensemble based approach, known as feature bagging for outlier detection (FBOD) [16], where at each iteration a subset of the feature space will be drawn randomly from a uniform distribution to determine the outlier scores for observations. After the completion of a pre-selected number of iterations, cumulative scores from these iterations will be counted as the outlier score for observations. In each iteration, a single outlier detection technique or multiple techniques can be utilized. In this work we use LOF as the only method and repeat it over iterations. We have already argued that any density and distance based method including the traditional LOF algorithm may suffer from the curse of dimensionality as it considers the entire feature space. By using LOF within FBOD is in a sense the combination of density and subspace based methods, which hopefully allows us to take full advantage of both.

IV. EXPERIMENTAL SETTINGS & RESULTS

For applying LOF method we need to select the value of k, which is the cardinality of MinPts. According to the suggestion of [6], we first set a lower bound and an upper bound for k. Then, we determine the outlier scores for every observation for all k values in between the bounds. For each observation, we choose the maximum of these scores as its final outlier score. The lower bound of k is chosen according to the suggestion in [6], which is 10, while the upper bound of k should be selected greater than the size of a hypothetical cluster of outliers that could form together. This can be done by observing the longest running streak of clustered events in a dataset. After consulting with the domain expert and data owner, this upper bound is set at 20. The first column of Table II contains the 30 top time stamps which have been identified as possible outliers from the LOF scores.

To apply the SOD method, we need to select the value of k at first and then based on that, the number of reference points. To maintain the comparability with LOF we choose $k = 15$, which is the average of the lower and upper bound of k used in the LOF method. Concerning a suitable number of reference points, it should be smaller than k but too small a value may render instability in computing the SOD scores. We explore a few options and finally settle on 10. Below 10, the SOD scores are not stable. It means that from the set of 15 nearest neighbors of any observation we have to select 10 of them as reference points which share the most similar neighbors if compared to the current observation's neighborhood. The second column of Table II contains the 30 top time stamps which have been identified as possible outliers from the SOD score.

In the FBOD method, we use LOF as our outlier detection technique and we run the algorithm for 50 iterations. In each iteration, a subset of the feature space has been selected. Finally, for each observation outlier scores are accumulated from each iteration. The third column of Table II contains the 30 top time stamps which have been identified as possible outliers from the FBOD score.

V. ANALYSIS OF RESULTS

The performance of the three methods are reasonably consistent as 22 out of top 30 time stamps returned by these methods as probable outliers are common (represented by dark blue color in Table II). This similarity extends to 33 if top 50 time stamps are considered and 55 if top 100 time stamps are considered (events beyond top 30 are not shown here to save space).

Another important insight is if we look closely at the results of our applied methods in Table II, we find that there are certain time chunks in a particular day (e.g. September 14, January 11, 12 etc.) which are more prone to outlier according to these methods. In some cases, specially when we move out from the range of top 30 time stamps, there are slight differences in the time stamps returned by individual methods but they were very close (within 10-50 mins range). The most possible explanation behind this phenomena is outliers appeared in a small cluster. We have

TABLE II
SUMMARY OF THE TOP 30 ANOMALIES RETURNED BY THE THREE METHODS. BLUE COLOR EVENTS ARE COMMON IN ALL COLUMNS.

LOF	SOD	FBOD
1/12/2016 11:30	9/14/2015 8:00	1/12/2016 11:30
9/14/2015 13:00	1/12/2016 11:30	9/14/2015 8:00
9/14/2015 13:10	9/13/2015 19:00	9/13/2015 19:00
1/12/2016 11:20	7/4/2015 8:30	1/9/2016 18:50
1/9/2016 18:50	7/4/2015 8:20	9/14/2015 13:00
1/2/2016 21:10	9/14/2015 1:50	9/14/2015 2:00
9/14/2015 8:00	7/4/2015 5:40	1/2/2016 21:10
1/2/2016 21:20	1/11/2016 12:00	9/14/2015 13:10
1/9/2016 18:30	9/14/2015 13:00	1/12/2016 11:20
9/14/2015 8:10	10/3/2015 14:40	1/11/2016 14:40
9/13/2015 19:00	7/4/2015 5:50	1/2/2016 21:20
9/14/2015 2:00	10/13/2015 8:15	9/14/2015 1:50
1/11/2016 14:40	9/14/2015 13:10	1/9/2016 18:30
1/11/2016 13:50	11/2/2015 9:56	9/14/2015 8:10
1/11/2016 12:00	7/4/2015 6:30	9/16/2015 10:50
1/11/2016 13:00	7/4/2015 4:30	1/9/2016 18:40
9/16/2015 10:50	1/2/2016 21:20	1/11/2016 12:00
9/17/2015 11:30	9/14/2015 2:00	10/3/2015 14:40
10/3/2015 14:40	9/14/2015 8:10	1/2/2016 21:30
1/2/2016 21:40	7/4/2015 4:20	1/9/2016 18:00
4/16/2015 23:10	1/11/2016 13:30	4/16/2015 23:10
10/4/2015 3:10	1/2/2016 21:40	4/16/2015 16:00
10/13/2015 8:15	7/4/2015 4:40	1/2/2016 21:40
10/14/2015 23:35	9/16/2015 10:50	11/2/2015 9:56
10/14/2015 23:15	1/2/2016 13:30	3/7/2016 9:40
1/2/2016 21:30	1/11/2016 14:40	10/4/2015 3:10
4/16/2015 16:00	1/2/2016 21:10	3/11/2016 12:30
11/2/2015 9:56	1/12/2016 11:20	10/13/2015 8:25
1/11/2016 13:30	1/9/2016 18:50	10/13/2015 8:15
9/14/2015 1:50	1/11/2016 13:00	1/11/2016 13:30

already discussed that the working mechanisms of LOF method and SOD method are completely different, whereas FBOD method lies somewhat in the middle. But in spite of their differences, they have returned similar results if we consider the top 30 time stamps. The significance is that using the multiple methods provide a strong cross validation among one another; otherwise, it is difficult to assess the validity of detection in an unsupervised learning circumstance.

The third observation is that the SOD method returns some of the time stamps from the 4th of July as outliers whereas the other two methods do not return any of the 4th of July time stamps as outliers. We know that, with the SOD method, it is possible to locate an outlier in a subspace consists of even only one variable, which is not possible in other two methods. So it is possible that those time stamps from the 4th of July showed significant deviation in any small dimensional subspace and thus not found by the other two methods. This presents an interesting question promoting the practitioners to look further into this specific cluster and decide if the variables in that small subspace are mis-handled (say, measurement errors) or if they are genuine outliers.

To find out which variables have more contribution to these outliers, we decide to select all the common events from the top 100 outliers identified by all three methods and deem them the true outliers. We chose the cut-off value as 100 after discussing with the domain expert. We find the number of common outliers out of this 100 to be 55. Assign the 55 events a response value of 1, and all other data records in the dataset a response value of 0 (meaning normal condition). As such, we convert the original unclassed data into a two-class dataset on which we build a classification and regression tree (CART) [17] using the R package `rpart` with their default parameter values; the resulting tree is shown in Fig. 3.

From this decision tree we can see that the variable *air pressure* and *delta oil temp - air temp of bearing F4* can correctly classify 24 of these 55 common outliers based on the right combination of their conditions. One such condition is when the air pressure is greater than 945 mbar and the difference between the oil temperature and the air temperature of bearing F4 becomes less than 10.582 degree Celsius, the generator almost surely behaves strangely, as the condition leads to nine anomalous observations consistently. This knowledge is useful for practitioners to do quick fault diagnosis and alarm management during operation.

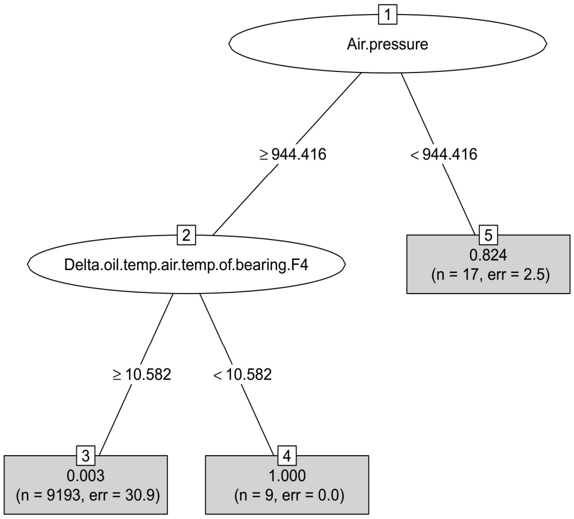

Fig. 3. Decision tree based on the top 55 anomalies found common from all three methods.

Another analysis that resulted in an important finding was to consider all the 100 anomalies found in the LOF model and apply a decision tree model. We found the threshold values for two important variables in the turbine system which include the oil temperature and harmonics values for bearing F4 as shown in Fig. 4. This was an important finding because during the preventive maintenance operation of the analyzed turbine, it was confirmed that bearing F4 needed repair to avoid future damage or costly interruption of the turbine operation to make the corrective action.

We have also tried to find out the important features

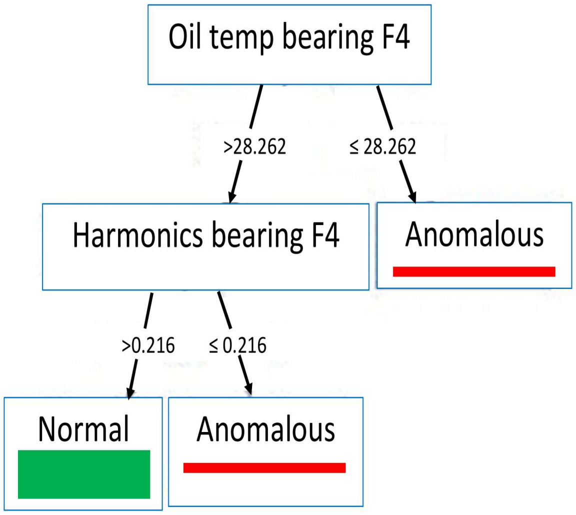

Fig. 4. Decision tree based on the 100 anomalies found from LOF method.

corresponding to the classification of outliers and normal observations using *Random Forest* based feature selection method [18]. We have used the R package `randomForest` with their default parameter values. According to our result, the significant features include *delta oil temp - air temp bearing F4*, *delta oil temp - air temp bearing F3*, *delta oil temp - air temp bearing F2*, *air pressure*, *oil temperature of bearing F2 and F4*, *metal temperature of bearing F2*, *delta metal temp - air temp bearing F2* and *harmonics of bearing F4*. These variables need to be monitored closely during the operation of the plant.

Furthermore, We trained a one-class SVM on the cleaner subset of the data after removing the 55 common outliers identified out of 100 in all three methods. For this, we used the R package `e1071` for training a SVM with radial basis kernel function. In this setting, we need to specify the values of one-class classification parameter, ν, and kernel function parameter, γ, which are selected based on a 10-fold cross validation as $\nu = 0.001$ and $\gamma = 0.01$. This resulting one-class SVM, when is applied to the whole dataset with the 55 outlier put back in, is able to separate all the presumed normal condition data and the 55 outliers successfully.

VI. Summary

A real life dataset received from a hydropower generation plant. After pre-processing, three anomaly detection methods are used to detect the anomalous observations which resulted in similar anomalies. The validity of the anomalies detected by the models is confirmed by the domain experts and maintenance operators. Although not yet observable in the physical system, the anomalies were observed during the preventive maintenance operation of the turbine. Root causes and threshold values for key attributes that contribute to the anomalies are determined in the form of decision tree. Critical variables leading to the anomalous observations are identified. This learned knowledge helps practitioners to monitor and diagnose the power plant during its operations. A one-class SVM classifier is trained and can be used to flag anomalies in future observations.

References

[1] A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," *Statistical Analysis and Data Mining*, vol. 5, no. 5, pp. 363–387, 2012.

[2] E. M. Knox and R. T. Ng, "Algorithms for mining distance based outliers in large datasets," in *Proceedings of the 24th International Conference on Very Large Data Bases*. Citeseer, 1998, pp. 392–403.

[3] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, vol. 29, no. 2. ACM, 2000, pp. 427–438.

[4] F. Angiulli and C. Pizzuti, "Outlier mining in large high-dimensional data sets," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 2, pp. 203–215, 2005.

[5] K. Zhang, M. Hutter, and H. Jin, "A new local distance-based outlier detection approach for scattered real-world data," in *Advances in Knowledge Discovery and Data Mining: Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2009, pp. 813–822.

[6] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, vol. 29, no. 2. ACM, 2000, pp. 93–104.

[7] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Loop: local outlier probabilities," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. ACM, 2009, pp. 1649–1652.

[8] E. Schubert, A. Zimek, and H.-P. Kriegel, "Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection," *Data Mining and Knowledge Discovery*, vol. 28, no. 1, pp. 190–237, 2014.

[9] L. J. Latecki, A. Lazarevic, and D. Pokrajac, "Outlier detection with kernel density functions," in *Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition*. Springer, 2007, pp. 61–75.

[10] H.-P. Kriegel, A. Zimek *et al.*, "Angle-based outlier detection in high-dimensional data," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. ACM, 2008, pp. 444–452.

[11] A. Zimek, R. J. Campello, and J. Sander, "Ensembles for unsupervised outlier detection: challenges and research questions a position paper," *ACM SIGKDD Explorations Newsletter*, vol. 15, no. 1, pp. 11–22, 2014.

[12] G. O. Campos, A. Zimek, J. Sander, R. J. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle, "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study," *Data Mining and Knowledge Discovery*, vol. 30, no. 4, pp. 891–927, 2016.

[13] J. Zhang, M. Lou, T. W. Ling, and H. Wang, "Hos (high dimensional outlying subspace)-miner: a system for detecting outlyting subspaces of high-dimensional data," in *Proceedings of the 30th International conference on Very Large Data Bases*. VLDB Endowment, 2004, pp. 1265–1268.

[14] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Outlier detection in axis-parallel subspaces of high dimensional data," in *"Advances in Knowledge Discovery and Data Mining: Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2009, pp. 831–838.

[15] E. Müller, I. Assent, U. Steinhausen, and T. Seidl, "Outrank: ranking outliers in high dimensional data," in *IEEE 24th International Conference on Data Engineering Workshop*. IEEE, 2008, pp. 600–603.

[16] A. Lazarevic and V. Kumar, "Feature bagging for outlier detection," in *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. ACM, 2005, pp. 157–166.

[17] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[18] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.