Anomaly Detection and Monitoring Building Infrastructure

XXXXXXX

07/09/2014

1 Monitoring Building Infrastructure

Large physical structures such as high-rise buildings, factories, and industrial complexes feature critical infrastructure that requires precise and ongoing monitoring. This infrastructure includes equipment such as elevators, heating and cooling systems, industrial machinery, electrical systems, security systems, and other components that need to be monitored to prevent outages, determine potential problems, and conserve resources. Monitoring equipment in these types of spaces has traditionally been very difficult for a variety of reasons. Many types of equipment provide sensor devices, and data provided by these devices is either not monitored at all – relying on the sensor itself instead to notify an individual, or the data is manually collected and somehow analyzed. Furthermore, equipment is often monitored independently, and rarely is monitoring data presented as a whole to building managers or other personnel.

Initial analysis indicates many of the problems encountered when monitoring building infrastructure are consistent with problems encountered with traditional data mining. Notable problems include collection of massive amounts of data, noise, inconsistent or incomplete data, pattern generation and mining, determining the interestingness of patterns, scalability, and efficiency. Building infrastructure data is usually collected from sensors, and is often related but heterogeneous. For example, replacement equipment may be installed for the same general purpose as preexisting equipment, but may be from different manufacturers and have different sensors

producing different data. The data may also contain noise or inconsistent values. Furthermore, hardware and sensor failures can cause missing or incomplete data sets. If not addressed correctly during data preprocessing, these various issues generally cause additional problems in subsequent stages, such as pattern generation and mining. Anomaly and outlier detection is a significant factor in dealing with these various problems, as it is used in both preprocessing and subsequent stages.

Anomaly and outlier detection is applied in different stages throughout the process of monitoring equipment data. First, anomaly detection can be applied during data preprocessing to smooth data. This stage in data preprocessing is generally referred to as noise reduction. Noise reduction can potentially have a negative impact when anomaly detection and problem identification are primary goals. If data points are eliminated in the noise reduction process, it may adversely affect our ability to identify problems, as the data points removed may be the same data that identifies a problem. Second, anomaly detection is applied in the pattern mining stage to identify problems. Anomaly detection at this stage can be challenging because outliers identified in relation to an overall dataset may not be problems in a smaller context. Conversely, problems identified in a smaller context may potentially need to be considered in relation to the overall dataset.

A combination of techniques can be used to balance the need to maintain the elimination of outliers as noise with the necessity of maintaining outliers for anomaly detection. Outliers have traditionally been identified as a result of cluster analysis or a clustering algorithm [1, Pg. 37], where the primary goal is to identify a data cluster, and outliers are simply a byproduct of non-association with any identified clusters. In these

scenarios, outliers are generally determined by evaluating their distance from a cluster, or conversely the cluster is calculated based on a distance threshold, and data points beyond the threshold are identified as outliers. The limitation in these scenarios is the primitiveness of the outlier classification – the data object is either an outlier or it is not an outlier. To begin to address some of the problems with noise reduction and anomaly detection, we should eliminate the notion of an outlier as a binary attribute, and instead assign an outlier factor or degree to a data object [2, Pg. 93]. Second, we can define thresholds that can be applied in both the noise reduction stage as well as the anomaly detection stage. This is a similar approach to the distance threshold used in cluster analysis, but it would be applied in multiple stages. A defined threshold in combination with an outlier degree can be used to identify which outliers are noise during preprocessing phases, and which are potential anomalies during pattern mining phases. Finally, we should also consider the neighborhood in which the object is located when considering it as an outlier, rather than the entire dimension of data. This will lead to an increase in accuracy when calculating an outlier's degree, because unrelated data in the dataset will not affect the data object's outlier factor.

A final issue to consider is efficiency of an anomaly detection process. Identifying problems is not helpful if the problems are identified too late. Many algorithms for outlier detection have been developed, and many perform well with small data sets, but do not scale up to large data sets. In addition, many algorithms do not perform well when dealing with high dimensional data. Scalability and efficiency issues present a significant problem when building infrastructure data needs to be monitored in real time. Scalability

and efficiency are a necessity when using anomaly detection to detect problems in large scale building infrastructures.

2 Anomaly Detection Overview

Breunig, Kriegel, Ng and Sander defined the concept of the *local outlier factor* (LOF) [2, Pg. 93]. The local outlier factor assigns a degree of being an outlier, rather than simply considering the object as an outlier or not. This concept is a departure from previous work in that it does not consider an object's outlier status as a binary attribute [2, Pg. 95]. This allows the use of thresholds, both for noise reduction, as well as anomaly detection. For example, a maximum threshold could be applied during the noise reduction phase, effectively eliminating objects with a degree exceeding the maximum threshold. A similar maximum threshold could be defined to detect anomalies. A maximum "problem" threshold can be identified, and therefore data objects exceeding these thresholds can be detected as potential problems. The "local" portion of the LOF refers to the object's outlier factor in relation to its local neighborhood, rather than the overall dataset [2, Pg. 94-95]. Evaluating data objects in relation to their local neighborhood is crucial because it can increase the accuracy of the object's outlier factor. This is an important notion when analyzing an overall dataset for a building's infrastructure for problems in specific areas or with specific equipment. It can help identify problems in relation to a specific portion or system of a building, rather than in relation to the building overall. Analysis of runtimes for algorithms using the local outlier factor in conjunction with various data sets indicate outliers can be accurately identified in near linear time [2, Figure 11].

Aggarwal and Yu introduced concepts for anomaly detection in high dimension data with many similarities to Breunig, et. al., but using an evolutionary search technique [1, Pg. 40]. The similarities included considering outliers as a byproduct of clustering algorithms, and outliers being evaluated in relation to a local neighborhood, rather than an overall dataset. Also, outliers were determined based on a "deviation value" [1, Pg. 37], similar to the local outlier factor. The usage of an evolutionary search technique introduced an additional concept: a higher level of interpretability and insight regarding why an object is an outlier [1, Pg. 39]. In terms of scalability and efficiency, the application of an evolutionary search technique proved to be better than other algorithms when processing large data sets [1, Pg. 44]. However, the proposed techniques continued to struggle when data was sparsely distributed throughout multiple dimensions [1, Pg.45]. This is an important problem when attempting to identify specific problem zones in the overall data space for a building's infrastructure.

Bay and Schwabacher proposed a promising algorithm which initial research indicates would scale well to large data sets [3, Pg. 32]. Their algorithm features nested loops, and has reduced run times from quadratic to linear time in various data sets. The algorithm is used in conjunction with randomization and pruning to achieve efficiency [3, Pg. 29]. However, this approach continues to present challenges. First, it assumes data is in random order. This can present a problem when monitoring building infrastructure data if data is dependent or related to a time element. A typical example is monitoring key card or security access. Determining problems with key card or security access in specific physical locations could be problematic if the time of attempted access is necessary and included as part of the anomaly detection algorithm. An additional

problem is the algorithm will perform poorly when the data does not contain outliers. The lack of outliers would defeat the purpose, as outliers are the very problems we are attempting to predict. Despite these various problems, the presented concepts are a promising contribution in regards to the scalability and efficiency problems that traditionally affect anomaly detection.

Bellala, Marwh, Shah, Arlitt, and Bash provide an approach for detecting anomalies in data generated by chiller equipment and cooling towers. Their approach applies a framework of finite state machine abstractions to assess patterns, detect anomalies, and predict potential problems [4, Pg. 153]. Their framework uses clustering and projection of data into low-dimension space, which improves the performance and scalability of the process. They also introduce the concept of thresholds for anomaly detection. Thresholds in this approach take the form of sustainability metrics, which are calculated based on data generated over time. The current state of data is then evaluated with these metrics to determine any anomalies [4, Pg. 154]. They also present an interesting concept regarding correcting missing values: a weighted global average is used to fill in the missing values caused by hardware or software failures [4, Pg. 156]. Therefore, instead of detecting and eliminating a missing value, the value is filled in and contributes to the accuracy of the overall data set. This can aid with anomaly detection, as it can increase the accuracy of a data object's outlier factor. They also evaluate anomalies in relation to theirs peers, which is similar to the neighborhood or local approach used by other algorithms. Finally, their framework is scalable to other datasets, and does not require an expert with domain knowledge.

3 Conclusion

Detecting and preventing problems in building infrastructure systems is important. Detecting and preventing problems can improve sustainability, reduce costs, improve operational continuity, and provide a myriad of other benefits. The challenges of collecting, consolidating, and analyzing datasets from infrastructure systems corresponds with challenges encountered in traditional data mining. More specifically, heterogeneous data sources, missing or incomplete data, pattern generation and mining, and pattern interestingness are all common problems encountered in data mining which also apply to monitoring building infrastructure.

Anomaly detection poses a unique challenge. Anomaly detection is often utilized in preprocessing stages to reduce noise. However, it can be further utilized to identify anomalies and potential problems in processed data. An important challenge in monitoring building infrastructure is applying anomaly detection appropriately to reduce noise, and subsequently using it to identify anomalies that might indicate a problem.

In noise reduction and subsequent anomaly detection stages, an outlier can be assigned a degree or factor of being an outlier, rather than simply being treated as an outlier or not. Thresholds can then be applied at both stages to then determine if an object is noise or represents a problem. The outlier factor can also be factored locally in relation to an object's neighborhood, rather than the overall dataset. This approach can increase the accuracy of an outlier's factor, and assist with identifying problems in specific sectors of data, which may translate to isolating problems within certain portion of a building's infrastructure.

Scalability and efficiency are crucial elements in anomaly detection. Identifying potential problems quickly can prevent them from occurring and causing subsequent problems. Various approaches to anomaly detection have displayed promising results when applied to data sets of various sizes, and ongoing research and analysis indicates algorithm performance will continue to improve.

4 References

1. Aggarwal C.C. and Yu P.S., "Outlier detection for high dimensional data," presented at the Proceedings of the 2001 ACM SIGMOD international conference on Management of data, Santa Barbara, California, USA, May 21-24, 2001.

2. Breunig M.M., Kriegel H., Ng R.T., and Sander J., "LOF: identifying density-based local outliers," presented at the Proceedings of the 2000 ACM SIGMOD international conference on Management of data, Dallas, Texas, USA, May 15-18, 2000.

3. Bay S. D., and Schwabacher M., "Mining distance-based outliers in near linear time with randomization and a simple pruning rule," presented at the Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, D.C., USA, August 24-27, 2003.

4. Bellala G., Marwah M., Shah A., Arlitt M., and Bash C., "A finite state machine-based characterization of building entities for monitoring and control," presented at the Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings, Toronto, Ontario, Canada, November 6, 2012.