Discovering Novelty in Time Series Data

Jonathan S. Anstey and Dennis K. Peters
Electrical and Computer Engineering
Memorial University of Newfoundland
St. John's, NL Canada A1B 3X5

Chris Dawson
INSTRUMAR Limited
St. John's, NL Canada A1B 4A5

Abstract

Modern manufacturing plants rely heavily on the use of automation. Automated facilities use sensors to measure material state and react to data patterns, which correspond to physical events. Many patterns can be predefined either by careful analysis or from domain experts. Instances of these patterns can then be discovered through techniques such as pattern recognition. However, this approach will fail to detect events that have not been predefined, potentially causing expensive production errors. A solution to this dilemma, novelty detection, allows for the identification of interesting data patterns embedded in otherwise normal data. This paper describes several novelty detection methods for time series data that have been proposed in the literature.

1 Introduction

The demand for improved efficiency and quality in the manufacturing industry has led to an increased use of automation. The success of automated production depends largely on control systems that monitor material state and take appropriate action. One method of monitoring uses sensors to detect changes in physical properties of the material being produced. In some cases, this measurement data has been interpreted manually by qualified human operators. This has worked well for many years in part because the amount of data has been small and manageable. Advances in data capture technology and the availability of inexpensive storage are making it possible to record and store vast amounts of measurement data. As a result, human operators are becoming less able to detect all material defects in a timely and cost effective fashion. There is a growing realization that for such monitoring systems to be effective, the data analysis must be automated.

Automating data analysis for the detection of faults is not a new concept in the manufacturing industry. For years computational intelligence techniques have been used to detect tool breakage, product defects, etc. [1, 2, 3]. On the forefront of industrial monitoring is INSTRUMAR Limited, a company based in Newfoundland and Labrador. This company has developed a sophisticated on-line synthetic polymer fiber monitoring system called the Attalus^{TM} Fiber System [4]. Synthetic polymer fiber is typically used in the clothing and textiles industries.

The basis for most of these techniques is to match the data to known patterns, which correspond to physical events. The matching process can be implemented with techniques such as pattern recognition, similarity search, or by other specialized means [3, 5]. Such systems are fine for detecting faults that have previously occurred. The procedure for doing so involves two steps: define data signatures for all possible faults, and then search for them continually. This works quite well if nothing unexpected happens. In this case, faults are expected patterns since we are searching for them. As shown in Figure 1, the obvious problem with this system is made evident when an unidentified event occurs. The system would fail to recognize the event and pass it off as normal operation data. In the case of a production line, this oversight could lead

to substantial financial losses in off specification product. In other words, sometimes the most interesting fault being searched for is not known. It may be unknown because it has never occurred before (as described above) or computing a data signature for it is too expensive [6].

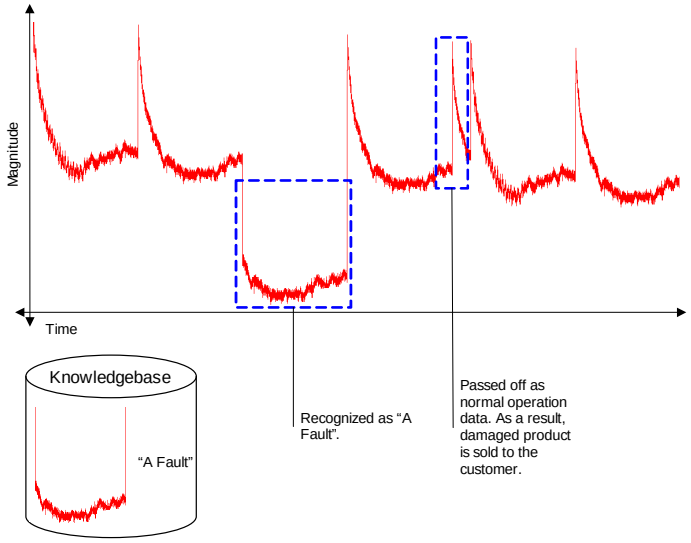

Figure 1: An incomplete knowledgebase will ultimately cause interesting events to be missed.

Clearly, enumerating all possible fault patterns can not detect all events. A more robust approach, called novelty detection, can be defined as the automatic identification of abnormal phenomena embedded in a large amount of normal data [7]. Much research has been devoted to this approach in general. Until recently however, novelty detection on time series data has received much less attention [8]. This is in part due to the fact that time series databases are usually very large and the notion of similarity can be subjective. Similarity may depend on the user, the domain, and the task at hand [9].

The occurrence of false positives is a problem for many novelty detection methods. A false positive occurs when a particular method marks something as novel when it is not. As a result, novelty detection systems should not be solely responsible for monitoring of a product. An ideal monitoring system could take advantage of both approaches. A pattern matching approach is very capable of finding known data patterns and novelty detection allows the system to expand its knowledgebase either autonomously or by a supervised means. It is the purpose of this paper to examine the various approaches to novelty detection on time series data that have been proposed in the literature.

2 Techniques

Many approaches to novelty detection on time series data have been proposed. Some mimic biological systems such as the human immune system [10] and the human brain [11]. Others apply classic statistical theory to the problem [12]. Still others approach the problem in a more unique way [8, 7, 13, 14].

2.1 Artificial Neural Networks

Artificial neural networks have been used for many years in the manufacturing industry for monitoring and control [15]. This is mainly because of their ability to learn patterns in data from experience - not from explicit mathematical models of the data. Neural networks are applied in cases where the underlying mathematical

models are too complex or too costly to determine by traditional means. They have been used successfully for novelty detection in time series data as well [11]. They learn/classify the normal data behavior and therefore distinguish normal from abnormal. For small problems neural networks work quite well. However, they do not scale well to massive datasets [8].

2.2 Negative Selection

Dasgupta and Forrest [10] proposed a novelty detection technique based on the negative selective mechanism of the human immune system. The human immune system is able to distinguish foreign (i.e. not seen before) cells from normal body cells. It does this by first generating a set of T-cells via a pseudo-random genetic rearrangement process. These T-cells are then exposed to the body's own cells in a training area. T-cells that bond to the body's cells are destroyed. The T-cells that are left match cells that are foreign to the body. This process is called negative selection [16].

The basic idea in applying this to data sets is to somehow generate a set of candidate pattern detectors and then discard those detectors that match the training data. The pattern detectors left over from this process are then used to match novel or abnormal patterns in the data. One major drawback of this method occurs when the normal data set becomes increasingly diverse. As a data set becomes more diverse, the number of possible patterns increases as well. As a result, more candidate detectors match the normal data and are destroyed. In the worst case, all pattern detectors are destroyed and the algorithm fails [8, 7].

2.3 Outlier Based

A novel occurrence can be thought of as being an outlier of normal occurrence data. Outlier analysis is well established in statistics as well as in the more recent data mining area. The authors of [12] introduce a scheme whereby optimal histograms are used to identify outliers in time series data. Under this scheme, outliers are defined as "*...points with values that differ greatly from that of surrounding points*" [12]. In a fault monitoring system a single data point does not give much information about the system state. System state is defined by multiple consecutive data points, a data sequence or pattern. We are most interested in data patterns that deviate from the normal operational data. As a result, this method as described is not suitable for fault monitoring. An interesting variation of this method involves searching for outliers in a feature space. A feature space is a transformed representation of the original data space. A simple example would be taking the standard deviation of the dataset at regular intervals. Each standard deviation value would be a representation of a sequence of values in the original data.

2.4 Wavelet Based

The authors of [13] take advantage of the multi-resolution property of the wavelet transform to find novel events at several levels of abstraction. For example, a high level of abstraction would ignore short novel events. Their approach uses a wavelet-based TSA-Tree (Trend and Surprise Abstractions Tree) structure to improve the performance of searching for novel events at multiple levels of abstraction. This method defines novelties as having a large difference between two consecutive averages [17]. This limits the detected novelties to being dramatic shifts in the signal. Because of this, short novelties that lie within the normal data pattern cannot be detected [8, 7]. Also, since the definition of surprise gives the same weight to positive and negative magnitude changes, there is no distinction made between them. Obviously for some systems, a high spike is not the same as a low spike. Again, the novelty definition imposes another flaw on this method. As described in [8], sometimes large changes in magnitude at regular intervals are part of the normal operational data. For instance, in a fiber production facility, the process of ending a package of fiber causes a large drop in the

signal at regular intervals. In this case the absence of a large drop in signal is a novel event. This algorithm would fail in this case.

2.5 Frequency Based

Keogh et al. [8] have developed a novelty detection algorithm based on data pattern frequencies. They define a novel pattern as having a frequency of occurrence different from that of normal data. Frequencies of all the normal data patterns are encoded in a suffix tree data structure. To utilize this structure, the data is first discretized into a textual representation. A Markov model is then used to estimate the expected frequency of any new data pattern. A data pattern with a greater frequency than what was expected is considered a novel event. As with any discretization technique, meaningful data may be lost in the transformation [7]. As a result, the discretization technique must be chosen carefully. An interesting property of this technique is its time/space complexity. Once the suffix tree is constructed from the normal data, the process of detecting all novel patterns in a dataset is linear in the size of the dataset. In 2003, a National Aeronautics and Space Administration (NASA) study singled out Keogh's approach as having "*great promise in the long term*" [18]. It is currently being tested by NASA for space shuttle launch monitoring [19].

2.6 Support Vector Machines

The authors of [7] propose a novelty detection method based on support vector machines (SVMs). SVMs are a relatively new machine learning technique applicable to both classification and regression. In this case, support vector regression (SVR) is applied. The idea behind SVR is to first map the input data via a nonlinear map to a higher dimensional feature space. By applying a simple linear algorithm in the feature space, a nonlinear fit is created in the input space. The use of kernel functions allows all computations to be carried out in the input space, with no explicit computations in the higher dimensional space. This makes SVR as powerful as a nonlinear method (such as neural networks) but much more computationally efficient [20].

Once the input data has been processed by the SVR method, all outliers of the model are considered to be novel events. An interesting feature of this method is its ability to associate a measure of confidence with each novel event. However, the authors admit that the technique has many open issues which require further investigation.

2.7 Rule Based

Another take on the novelty detection problem is to learn rules from the data that define the possible normal sequences [14]. Any sequence of data deviating from these rules would be considered a novelty. This approach involves three main steps as illustrated in Figure 2 (adapted from [14]). The first step is to separate the raw data into clusters or segments. Next, rules that best describe each segment are generated. In the illustration, signal magnitude and slope are the defining features of each segment; other measures may be used however. From these rules, a state machine can be constructed that will only accept normal data sequences. Any sequence that is not accepted by the state machine is marked as novel. This method was designed to automate the process of populating an expert system's knowledgebase. As a result, the rules generated by this method are human readable - a contrast to the other novelty detection methods described above. Having the ability to clearly explain and tweak the underlying behavior is of great value.

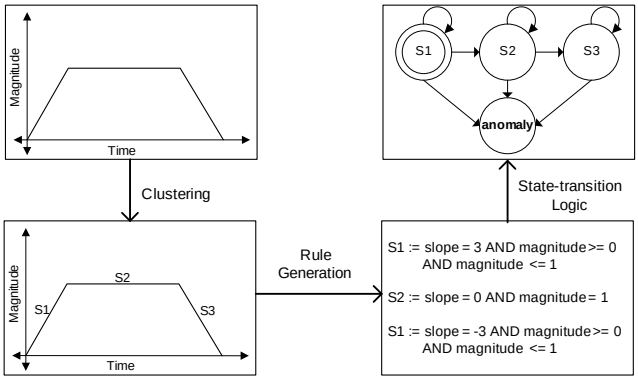

Figure 2: Learning rules to uncover novelties

3 Conclusions and Future Work

As modern manufacturing facilities grow more automated, product quality is depending more on automation systems. The most effective automation solution detects all possible events. Novelty detection is an integral part of such a solution. It is currently a "hot spot" for research with many different approaches being investigated in parallel. Currently, it is not clear as to which method is most suitable to a fiber monitoring application.

Future work will focus on evaluating the effectiveness of several of the aforementioned novelty detection techniques to the problem of identifying faults in real fiber sensor data. This will involve creating a software prototype for each technique and evaluating them with real fiber data. The prototypes will be compared based on two metrics: probability of detection and probability of false alarms. Probability of detection refers to the rate that a method will successfully detect a novel data pattern during a particular time frame. Probability of false alarms refers to the rate that a method will detect data patterns that are not particularly novel. In addition, a recommendation will be made on the method that is best suited to applications involving fiber.

4 Acknowledgment

The authors gratefully acknowledge the funding support provided by INSTRUMAR Limited and the Natural Sciences and Engineering Research Council of Canada, through the Industrial Postgraduate Scholarship.

References

[1] K. Goebel, B. Wood, A. Agogino, and P. Jain, "Comparing a neural-fuzzy scheme with a probabilistic neural network for applications to monitoring in manufacturing systems," in *Working Notes of the 1994 AAAI Spring Symposium: Detecting and Resolving Errors in Manufacturing Systems*, 1994.

[2] X. Li, "Fuzzy neural network and wavelet for tool condition monitoring," in *Computational Intelligence in Manufacturing Handbook*, CRC Press, 2001.

[3] J. Dorrity, G. Vachtsevanos, and W. Jasper, "Real-time fabric defect detection and control in weaving processes," in *National Textile Center Annual Report*, pp. 113–122, 1996.

[4] M. Chan, "New online fiber sensor technology unlocks value in fiber manufacturing," in *International Fiber Journal*, December 2000.

[5] R. Agrawal, K.-I. Lin, H. S. Sawhney, and K. Shim, "Fast similarity search in the presence of noise, scaling, and translation in time-series databases," in *VLDB '95: Proceedings of the 21th International Conference on Very Large Data Bases*, (San Francisco, CA, USA), pp. 490–501, Morgan Kaufmann Publishers Inc., 1995.

[6] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, 2004.

[7] J. Ma and S. Perkins, "Online novelty detection on temporal sequences," in *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, (New York, NY, USA), pp. 613–618, ACM Press, 2003.

[8] E. Keogh, S. Lonardi, and B. Chiu, "Finding surprising patterns in a time series database in linear time and space," in *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, (New York, NY, USA), pp. 550–556, ACM Press, 2002.

[9] E. Keogh, "Mining and indexing time series data," in *The 2001 IEEE International Conference on Data Mining*, (San Jose, CA, USA), 2001.

[10] D. Dasgupta and S. Forrest, "Novelty detection in time series data using ideas from immunology," in *Proceedings of the 5th International Conference on Intelligent Systems*, 1996.

[11] A. L. I. Oliveira, F. B. L. Neto, and S. R. de Lemos Meira, "Combining MLP and RBF neural networks for novelty detection in short time series.," in *MICAI*, pp. 844–853, 2004.

[12] H. V. Jagadish, N. Koudas, and S. Muthukrishnan, "Mining deviants in a time series database," in *VLDB '99: Proceedings of the 25th International Conference on Very Large Data Bases*, (San Francisco, CA, USA), pp. 102–113, Morgan Kaufmann Publishers Inc., 1999.

[13] C. Shahabi, X. Tian, and W. Zhao, "TSA-tree: A wavelet-based approach to improve the efficiency of multi-level surprise and trend queries on time-series data," in *SSDBM '00: Proceedings of the 12th International Conference on Scientific and Statistical Database Management (SSDBM'00)*, (Washington, DC, USA), p. 55, IEEE Computer Society, 2000.

[14] S. Salvador, P. Chan, and J. Brodie, "Learning states and rules for time series anomaly detection," in *Proceedings of the 17th international FLAIRS conference*, pp. 300–305, 2004.

[15] J. Wang, W. S. Tang, and C. Roze, "Neural network applications in intelligent manufacturing: An updated survey," in *Computational Intelligence in Manufacturing Handbook*, CRC Press, 2001.

[16] D. Dasgupta, Z. Ji, and F. Gonzlez, "Artificial immune system (AIS) research in the last five years," in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*, (Canberra, Australia), 2003.

[17] Y. Zhu, *High Performance Data Mining in Time Series: Techniques and Case Studies*. PhD thesis, New York University, New York, 2004.

[18] D. Isaac and C. Lynnes, "Automated data quality assessment in the intelligent archive." White Paper Prepared for the Intelligent Data Understanding Program, January 2003.

[19] J. Lin, E. Keogh, S. Lonardi, J. P. Lankford, and D. M. Nystrom, "Visually mining and monitoring massive time series," in *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, (New York, NY, USA), pp. 460–469, ACM Press, 2004.

[20] M. A. Hearst, "Support vector machines," *IEEE Intelligent Systems*, vol. 13, no. 4, pp. 18–28, 1998.