Analysis of Time Series Mining in Manufacturing Problems

Ruhaizan Ismail, University Kebangsaan Malaysia, Malaysia
Zalinda Othman, University Kebangsaan Malaysia, Malaysia
Azuraliza Abu Bakar, University Kebangsaan Malaysia, Malaysia

The Asian Conference on Technology, Information & Society 2014
Official Conference Proceedings

Abstract

Recently, almost all modern manufacturing processes use operation that rely on automatic tools. The automatic tools react by material used and physical events. Physical events are determined by data patterns that can be definable by domain experts. However, when problems have not been predefined, it could cause critical errors especially in manufacturing plants. Furthermore, with machine complexity and new technology, this leads to a huge amount of data. These problems call for mining extraction to useful knowledge or patterns. Time series mining is about solving these problems by applying various approaches that suitable with manufacturing data. Until recently, the detection of time series data has received much less attention because of time series databases are usually very large, high dimensionality and the concepts of similarity can be subjective. The similarity may depends on the user, the domain, and the task at hand. As a result, a suitable time series mining approaches can be used to find any interesting and useful patterns. This study intents to present a broad classification of various time series problems in manufacturing operations. Also, the theoretical developments and analysis of the current available approaches used in time series mining are reviewed.

Keywords: data mining, time series, manufacturing, review

iafor
The International Academic Forum
www.iafor.org

Introduction

Time series data appears naturally in almost fields of natural and social science as well as in numerous other disciplines. Most people are familiar with financial time series data like daily stock prices. For economist, they may want to identify the trend of changes in an annual household income over time and the relationship between different time series such as annual household incomes from different regions. In meteorological research, time series data can be mined for predictive analysis such as monthly average temperature (Zyu, 2004). Time series data also contain a valuable hidden knowledge within design and manufacturing environment. Time series databases consist of sequences of values or events obtained over repeated measurements of time. The values are typically measured at equal time intervals (e.g., hourly, daily, weekly). Thus, database contains time series data is called a sequence database. In a sequence database, each data is associated with the sequence of ordered events, with or without concrete notions of time (Han & Kamber, 2006).

This paper reviews on time series mining in manufacturing problems that apply various suitable approaches to manufacturing data. Until recently, time series mining has received much less attention because of time series databases are usually very large, high dimensionality and the concept of similarity can be subjective. The similarity may depends on the user, the domain, and the task at hand. As a result, a suitable time series mining can be used to find an interesting and useful data patterns in manufacturing operation. This study intents to present a broad classification of various time series manufacturing problems.

1. Time series mining

Temporal databases capture time-related attributes whose value change with time, for example, stock exchange data. There are five types of temporal data, which are static, sequences, time stamped, time series and fully temporal. However, two types of temporal data are dominant in the development of temporal data mining, i.e time-series data and sequence data. Time series data is defined as a time-ordered sequence of observation values of a physical or financial variables made at equally spaced time intervals (Palit & Popovic, 2006). It also can be a stretch of values on the same scale indexed by a time-like parameter, such as a range over the positive and negative integers or all real number or subsets of these. Also, the data is defined as a sequence of a real number that varies with time like stock prices, exchange rates, biomedical measurements data and so forth (Hsu et al., 2008). Another definition from Fu (2011), time series data is a collection of observations made by chronologically.

For sequence data, it is defined as a list of transaction, and transaction time is associated with each transaction (Hsu et al., 2008), or a sequence is an ordered list of event. An event can be represented as a symbolic value, a numerical, a vector or a complex data type. According to Xing et al. (2010), sequence data has subtypes that consists of a simple symbolic sequence, a complex symbolic sequence, a simple time series and a multivariate time series. A simple symbolic sequence is an ordered list of the symbols from the alphabet. However, the sequence like DNA sequence or protein sequence is called a complex symbolic sequence. It is complex because it is an ordered list of vector. Meanwhile a simple time series is a sequence of real values in time stamp from t_1 to t_n. However, when the simple time series used, the data needs to be transformed into symbolic sequence through discretization or symbolic transformation. Meanwhile, a multivariate time series is a sequence of numerical vectors. It has been used for gesture and motion recognition (Xing et al., 2010).

Brillinger (2000) stated there are two categories of problem arise in time series data. There are scientific and statistical problem. The scientific problem is a problem that related with a systematic approach. It is designed to eliminate bias in the search of validation of facts. An example of scientific problems are measurement of uncertainty, difficulty in constructing a relationship, and spatially aggregated data (Brillinger, 1994). The systematic approaches that deal with scientific problems are smoothing, prediction, and association**.** However, more researches focus on solving problems related to predictive and associative approaches. This because of its related to forecasting the future values. Whereas, the statistical problem is a problem that associated with facts gathering and data analysis related to the process as a whole. There are example problems such as hidden frequencies, uncertainty computation, goodness of fit and testing. Beside these two problems, there are also special difficulties arise, such as missing values, censoring, measurement error, irregular sampling, feedback, outliers, shocks, signal-generated noise, trading days, festivals, changing seasonal pattern, measurement error, aliasing, and data observed in two series at different time points.

From the database perspective, time series data are facing problems such as large in data size, high dimensionality dan necessary to update continuously (Fu, 2011). In manufacturing, the large amount of data happens when the data is generated and collected during hourly or daily operations. Data through time, process and event are collected and stored in the database at various stages of design and production (Braha, 2001). This data may relate with many parameters such as products, materials, design and processes. The data contain hundreds of attributes, and the domain for each attribute can be large. This enormity is called high dimensionality data (Agrawal et al., 1998). From Esling & Agon (2012), time series data is essentially a high dimensional data. The high dimensionality problem makes the data difficult to be represented, imbalance proportion, long time scale and hinder in useful knowledge extraction. Because of the continuously updated characteristic, time series data also being called as dynamic or transactional data and it difficults to be captured. This nature makes the time series data mining more challenging. Time series researches are categorized into few different areas as:

i. Representation and indexing
   This task aims for dimensionality reduction. There are several methods in this task that representing time and frequency domains. In time domain approaches, the simplest and earlier developed method is sampling (Astrom, 1969). From sampling, the enhanced methods are developed by applying the mean value, approximating straight line, preserving the essential points or converting numeric time series to symbolic form. Representing time series in a frequency domain also a popular transformation technique. The proposed techniques are such as discrete Fourier transforms (Agrawal et al., 1993), similarity-based queries (Rafiei & Mendelzon, 2000), likelihood ratio statistics (Janacek et al., 2005) and discrete wavelet transforms (Chan & Fu, 1999). Other representing time series methods are hidden Markov models (Azzouzi & Nabney, 1998) and multidimensional indexing structure.

ii. Similarity measure

Similarity measure is like an exact match that based on the simple database definition. In similarity search, it finds data sequences that slightly differ from the given query sequence. Using an index to retrieve the sequence the actual distance between sequences is computed and discard any false matches. With the concept of matching, the similarity measure is represented in two ways, i.e. a whole sequence matching and subsequence matching. In the whole sequence matching, the query sequence is compared to each candidate in the series with the smallest distance. Example of the approaches are Euclidean distance, a constraint-based similarity query (Goldin & Kanellakis, 1995), pattern recognition (Morrill, 1998), dynamic time warping distance measure (Berndt & Clifford, 1994), threshold-based distance function (Abfalg et al., 2006) and parameter-light distance measure (Keogh et al., 2007). For subsequence matching, it's comparing the query sequence with the subsequence in the longer time series. The query sequence is required to be placed at every offset within the longer time series. The proposed subsequence approaches are window ordering (Kim & Jeong, 2007), linear segments (Morinaka et al., 2001), dynamic time warping with a suffix tree (Gusfield, 1997), and minimum-distance matching-window pair (Han et al., 2007). Also, a graph, framework or language can represent the relationship between subsequences and pattern query.

iii. Segmentation

Segmentation is a practice of dividing knowledge or information into groups that are similar in specific ways. It is considered as a discretization problem in a processing step. Earlier, a fixed length window has been proposed in order to segment data into a subsequence. There are two major disadvantages in using this approach, where first, the time series appeared meaningful pattern in different lengths, and second, meaningful pattern may be lost during cutting points. In order to prevent these problems, segmentation is handled by using a dynamic approach which identifies time points in more flexible way, i.e. the concept of different window widths. From that, more techniques have been proposed such as fuzzy clustering based segmentation (Abonyi et al., 2003), piecewise generalized likelihood ratio (Wang & Willett, 2004), and sliding test window (Chu, 1995). On the other hand, another way of solving the segmentation problem is by finding cyclic periodicity for all of the segments.

iv. Visualization

Visualization is one of the essential tools/ways in order to present the processed time series for further analysis. Many tools have been developed by past researchers such as by using pattern-based analysis (Schreck et al., 2007), and querying (Keogh et al., 2002). Also, other visualization techniques have been proposed such as a tree i.e. by converting each subsequence into a symbol string (Lin et al., 2005), bitmap (Kumar et al., 2005) and dot plots (Yankov et al., 2005).

v. Mining

Essentially, mining is aimed to discover hidden information or knowledge from either the original or transformed data. There are four issues arise in mining; pattern discovery, classification, rule discovery and summarization.

Pattern discovery is one of data mining task that mining various kinds of pattern, sequential patterns or sub-graph patterns. Another task in data mining is classification. Classification in time series mining usually transformed time series data into sequence data. From that, the task used to generate classifiers to find patterns and mining time series data into a useful result. For rule discovery, it is a task that mainly focus on symbolic items present in transactions. Rules are generated to describe each segment and in the terms of human readable. The advantages of rules, it can be explained clearly and tweak the underlying behavior. In summarization task, it is more on producing a compact description based on short or long-term time series data collection.

2. Issues in manufacturing

In most sectors, manufacturing is extremely competitive, and the financial margins that differentiate between success and failure are very tight, with most established industries needing to compete, produce and sell at a global level. Master to these trans-continental challenges, a company must achieve low-cost production. Still, it can maintain highly skilled, flexible and efficient workforces who can consistently design and produce high quality and low products (Choudhary et al., 2008).

There are various researches that review and focus on data mining in different application areas of manufacturing. The areas are design engineering, manufacturing systems, decision support system, shop floor control and layout, fault detection and quality improvement, maintenance, and customer relationship management. Choudhary et al. (2008), reviewed the various past researchers for data mining in manufacturing using five data mining functions; concept description, classification, clustering, prediction and association. They also categorized these functions for the manufacturing domains into quality control, job shop scheduling, fault diagnosis, manufacturing process, manufacturing system, maintenance, defect analysis, etc. Meanwhile, in Shanawaz et al. (2011), they presented an overview of eight techniques of temporal data mining; association, prediction, classification, clustering, characterization, search and retrieval, pattern discovery and trend analysis. As for Fu (2011), a review about time series in data mining, which was characterized into five tasks that already elaborate in section 2; representation and indexing, similarity measure, segmentation, visualization and mining. With focusing only in mining tasks, this paper highlights the past researchers in the manufacturing domain with pattern discovery issues.

From Fu (2011), pattern discovery is one of the issues that involves finding existing and surprising patterns. Pattern discovery also known as motif discovery, anomaly detection or finding discords. Pattern mining usually involved with temporal data because it implies the existence of time. There are three types of analysis that used in temporal data, such as temporal association rules, sequential pattern, and periodic patterns (Hsu et al., 2008). Temporal association rules regularly sampled univariate time series data. However, sequential pattern usually sampled in multivariate time series data. Sequential pattern concerned with finding precedence relationships. This pattern ordered association among data examples in the sequence. Some commonly used algorithms are Apriori, GSP, PrefixSpan and SPADE. Whereas, periodic pattern is considered as temporal regularity. This is usually used in web usage recommendation, weather prediction, computer networks and biological data (Huang & Chang, 2004).

With a focus on pattern discovery approaches, past researchers have reviewed several common problems in manufacturing operations.

Database/product fault

Zaki et al. (2000) presented an algorithm to extract patterns of events that predict failures in databases of plan executions. Analyzing execution traces is appropriate for planning that contain uncertainty, such as incomplete knowledge of the world or action with probabilistic effects. The causes of plan failures were extracted to feed the discovered patterns back into the planner. The goal is to find "interesting" sequences that have a high confidence of predicting plan failure. SPADE was used to mine such patterns.

Fountain et al. (2003), described a decision analysis for problem in integrated circuits manufacturing. The purpose of the testing was to avoid the expenses of packaging bad die and provide feedback by detecting die failures. They used a decision-theoretic approach to create a probabilistic pattern model of die failures. This model has been combined with a computation in deciding which die to test next and when to stop testing in real time.

Buddhakulsomsiri & Zakarian (2009) presented a sequential pattern mining algorithm that allow product and quality engineers to extract hidden knowledge from a large automotive warranty database. The algorithm used the elementary set concept and database manipulation techniques to search for patterns or relationships among occurrences of warranty claims over time. The sequential patterns represented in the form of association rules. The generated rules include the quality or warranty problems, and labor codes that occurred at a later time. Once a set of unique sequential patterns was generated, the algorithm applied a set of thresholds to evaluate the significance of the rules. The significant of rules provided the knowledge about how many product failures that lead to the future product faults

Maintenance

Sodiya et al. (2005) implemented Data mining-based Intelligent Maintenance System (DIMS) algorithm using Visual Basic 6.0 (VB). The algorithm applied a modified Apriori algorithm for frequent pattern mining on database. It used to extract predictive information needed for maintenance purposes and to identify association rules among faults. This system was designed to provide a way for timely and sophisticated analysis of fault database for effective equipment maintenance.

Quality control

Research by Purintrapiban and Kachitvichy (2003) proposed a fractal dimension based classifier in quality control. The proposed classifier was used to detect any unnatural patterns in the process. Fractal dimension is a recent classifier for pattern classification. The classifier is an index for measuring the complexity of an object. The results showed the classifier was effective in detecting non-periodic patterns such as natural patterns, linear patterns, systematic variable, stratification, mixture and sudden shifts. However, this approach was not sensitive for detecting linear patterns with a small slope.

Knowledge acquisition

Irani et al. (1993) developed an expert system for predicting the result of future experiments under various conditions, i.e. noisy data and limited availability training data. The system used generalized ID3 algorithm to diagnose and optimize a reactive ion etching process. The expert system was built for a knowledge acquisition. The discovered pattern was consistent with the data and has the same expectation from expert.

In 2003, Woon et al. proposed a method called PDMiner to mine web logs efficiently. This method was developed by using tree structure, association rule mining and sequential pattern mining techniques. PDMiner discovers the relationship among parts, assemblies, documents and how people interact with one another.

3. Our current research

From Choudhary et al. (2008), association rule and clustering have been proven as an effective method for extracting common patterns. These two approaches have mostly been used in the manufacturing field. However, there are a lot of problems in manufacturing industry that cannot be solved with only these two methods. The challenges are harder when it involves classifying sequence data or extracting pattern or knowledge. Time series methods have the potential to process and mine complex manufacturing data.

In our research, one framework are proposed for knowledge extraction from production data. A predictive model will be build from the extracted patterns by using sequential pattern approaches. The model can be used for identifying and characterizing sequence families, mining order from sequences and distinguishing any interesting sequences. Algorithms from sequential pattern mining are suitable for a large dataset and greatly reduced the amount of subsequence candidates compared to other approaches. The algorithms extract any hidden knowledge using pseudo-projection from frequent items within the data. The hidden knowledge of total execution time, lead time and makespan can be used to imply machine utilization and workload. This knowledge can be used to minimize capacity loss during the planning or rescheduling process.

4. Conclusions

From this study, a suitable time series mining approach can be used to find any interesting and useful data patterns in manufacturing operations. We have presented an overview of techniques for time series in time series mining. With many interesting techniques of solving time series data, it has been shown to be useful in many applications, especially in manufacturing domain. Past researchers have been tried to build a knowledge model for time series data. If a model is successful in interpreting the observed time series, the future value can be predicted to hold in the future.

Acknowledgements

The first author acknowledges the financial support provided under grant FRGS/1/2014/ICT02/UKM/02/1 Fakulti Teknologi Dan Sains Maklumat, Universiti Kebangsaan Malaysia.

References

Aßfalg, J., Kriegel, H.-P., Kröger, P., Kunath, P., Pryakhin, A., & Renz, M. (2006). Similarity search on time series based on threshold queries. *Advances in Database Technology-EDBT 2006*, 276-294: Springer.

Abonyi, J., Feil, B., Nemeth, S. & Arva, P. (2003). Principal Component Analysis Based Time Series Segmentation-Application to Hierarchical Clustering for Multivariate Process Data. *IEEE international conference on computational cybernetics,* 29-31.

Agrawal, R., Faloutsos, C. & Swami, A. (1993). Efficient Similarity Search in Sequence Databases. Springer.

Agrawal, R., Gehrke, J., Gunopulos, D. & Raghavan, P. (1998). Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. ACM.

Åström, K. J. (1969). On the Choice of Sampling Rates in Parametric Identification of Time Series. *Information Sciences*, *1(3),* 273-278.

Azzouzi, M. & Nabney, I. T. (1998). Analysing Time Series Structure with Hidden Markov Models. *Proceedings of the IEEE Conference on Neural Networks and Signal Processing*, 402-408.

Berndt, D. J. & Clifford, J. (1994). Using Dynamic Time Warping to Find Patterns in Time Series. *KDD workshop*, 359-370.

Braha, D. (2001). Data Mining for Design and Manufacturing - Methods and Applications: Kluwer Academic Publisher.

Brillinger, D. R. (1994). Examples of Scientific Problems and Data Analyses in Demography, Neurophysiology, and Seismology. *Journal of Computational and Graphical Statistics, 3(1)*, 1-22.

Brillinger, D. R. (2000). Time Series: General. Int. Encyc. Social and Behavioral Sciences.

Buddhakulsomsiri, J. & Zakarian, A. (2009). Sequential Pattern Mining Algorithm for Automotive Warranty Data. *Computer and Industrial Engineering, 57(1).*

Chan, K.-P. & Fu, A.-C. (1999). Efficient Time Series Matching by Wavelets. Data Engineering, 1999. *Proceedings., 15th International Conference on*, 126-133.

Choudhary, A. K., Harding, J. A. & Tiwari, M. K. (2008). Data Mining in Manufacturing: A Review Based on the Kind of Knowledge. *Journal Intelligent Manufacturing, 20(5)*, 501-521.

Chu, C.-S. J. (1995). Time Series Segmentation: A Sliding Window Approach. *Information Sciences, 85(1),* 147-173.

Esling, P. & Agon, C. (2012). Time-Series Data Mining. *ACM Computing Surveys (CSUR), 45(1),* 12.

Fountain, T., Dietterich, T. & Sudyka, B. (2003). Data Mining for Manufacturing Control: An Application in Optimizing Ic Tests. *Exploring artificial intelligence in the new millennium,* 381-400.

Fu, T.-C. (2011). A Review on Time Series Data Mining. *Engineering Applications of Artificial Intelligence, 24(1)*, 164-181.

Goldin, D. Q. & Kanellakis, P. C. (1995). On Similarity Queries for Time-Series Data: Constraint Specification and Implementation. *Principles and Practice of Constraint Programming—CP'95*, 137-153.

Gusfield, D. (1997). Algorithms on Strings, Trees and Sequences. *Computer Science and Computational Biology*, Cambridge University Press.

Han, J. & Kamber, M. (2006). Data Mining Concepts and Techniques. Academic press. Morgan Kaufmann.

Han, W.-S., Lee, J., Moon, Y.-S. & Jiang, H. (2007). Ranked Subsequence Matching in Time-Series Databases. *Proceedings of the 33rd international conference on Very large data bases*, 423-434.

Hsu, W., Lee, M. H. & Wang, J. (2008). Temporal and Spatio-Temporal Data Mining. New York, IGI Publishing.

Huang, K.-Y., & Chang, C.-H. (2004). Mining periodic patterns in sequence data. *Data Warehousing and Knowledge Discovery*, 401-410: Springer.

Irani, K. B., Cheng, J., Fayyad, U. M. & Qian, Z. (1993). Applying Machine Learning to Semiconductor Manufacturing. *IEEE Expert, 8(1),* 41-47.

Janacek, G.J., Bagnall, A.J., & Powell, M. (2005). A likelihood ratio distance measure for the similarity between the fourier transform of time series. *Advances in Knowledge Discovery and Data Mining*, 737-743: Springer.

Keogh, E., Hochheiser, H., & Shneiderman, B. (2002). An augmented visual query mechanism for finding patterns in time series data. *Flexible Query Answering Systems*, 240-250: Springer.

Keogh, E., Lonardi, S., Ratanamahatana, C. A., Wei, L., Lee, S.-H. & Handley, J. (2007). Compression-Based Data Mining of Sequential Data. *Data Mining and knowledge discovery, 14(1),* 99-129.

Kim, S.-W. & Jeong, B.-S. (2007). Performance Bottleneck of Subsequence Matching in Time-Series Databases: Observation, Solution, and Performance Evaluation. *Information Sciences, 177(22),* 4841-4858.

Kumar, N., Lolla, V. N., Keogh, E. J., Lonardi, S. & Ratanamahatana, C. (2005). Time-Series Bitmaps: A Practical Visualization Tool for Working with Large Time Series Databases. *SDM*, 531-535.

Lin, J., Keogh, E. & Lonardi, S. (2005). Visualizing and Discovering Non-Trivial Patterns in Large Time Series Databases. *Information visualization, 4(2),* 61-82.

Morinaka, Y., Yoshikawa, M., Amagasa, T. & Uemura, S. (2001). The L-Index: An Indexing Structure for Efficient Subsequence Matching in Time Sequence Databases. *Proc. 5th PacificAisa Conf. on Knowledge Discovery and Data Mining*, 51-60.

Morrill, J. P. (1998). Distributed Recognition of Patterns in Time Series Data. *Communications of the ACM, 41(5),* 45-51.

Palit, A. K. & Popovic, D. (2006). Computational Intelligence in Time Series Forecasting. *Theory and Engineering Applications*, Springer.

Purintrapiban, U. & Kachitvichyanukul, V. (2003). Detecting Patterns in Process Data with Fractal Dimension. *Computers & industrial engineering, 45(4),* 653-667.

Rafiei, D. & Mendelzon, A. O. (2000). Querying Time Series Data Based on Similarity. *Knowledge and Data Engineering, IEEE Transactions, 12(5),* 675-693.

Schreck, T., Tekušová, T., Kohlhammer, J. & Fellner, D. (2007). Trajectory-Based Visual Analysis of Large Financial Time Series Data. *ACM SIGKDD Explorations Newsletter, 9(2),* 30-37.

Shanawaz, M., Ranjan, A. & Danish, M. (2011). Temporal Data Mining: An Overview. *International Journal of Engineering and Advanced Technology (IJEAT), 1(1),* 20-24.

Sodiya, A., Longe, H. & Ibrahim, S. (2005). Data Mining-Based Intelligent Equipment Maintenance System in Telecommunication Network. *Journal of Applied Computer Science, 13(2),* 29-38.

Wang, Z. J. & Willett, P. (2004). Joint Segmentation and Classification of Time Series Using Class-Specific Features. Systems, Man, and Cybernetics, Part B: Cybernetics. *IEEE Transactions, 34(2),* 1056-1067.

Woon, Y.-K., Ng, W.-K., Li, X. & Lu, W.-F. (2003). Efficient Web Log Mining for Product Development. Cyberworlds, 2003. *Proceedings. 2003 International Conference on*, 294-301.

Xing, Z., Pei, J. & Keogh, E. (2010). A Brief Survey on Sequence Classification. *ACM SIGKDD Explorations Newsletter, 12(1),* 40-48.

Yankov, D., Keogh, E. & Lonardi, S. (2005). Dot Plots for Time Series Analysis. Tools with Artificial Intelligence, ICTAI 05. *17th IEEE International Conference*,168.

Zaki, M. J., Lesh, N. & Ogihara, M. (2000). Planmine: Predicting Plan Failures Using Sequence Mining. *Artificial Intelligence Review, 14(6),* 421-446.

Zhu, Y. (2004). High Performance Data Mining in Time Series: Techniques and Case Studies. *Tesis Degree of doctor of Philosophy, Department of Computer science, New York University*.