Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 61 (2015) 395 – 401
Procedia Computer Science

Complex Adaptive Systems, Publication 5
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2015-San Jose, CA

Determination of Rule Patterns in Complex Event Processing Using Machine Learning Techniques

Nijat Mehdiyev[a], Julian Krumeich[a], David Enke[b*], Dirk Werth[a], Peter Loos[a]

[a]*Institute for Information Systems (IWi). German Research Center for Artificial Intelligence (DFKI), 66123 Saarbrücken, Germany*
[b]*Department of Engineering Management and System Engineering, Missouri University of Science and Technology, Rolla, MO, 65409-0370, USA*
Laboratory for Investment and Financial Engineering

Abstract

Complex Event Processing (CEP) is a novel and promising methodology that enables the real-time analysis of stream event data. The main purpose of CEP is detection of the complex event patterns from the atomic and semantically low-level events such as sensor, log, or RFID data. Determination of the rule patterns for matching these simple events based on the temporal, semantic, or spatial correlations is the central task of CEP systems. In the current design of the CEP systems, experts provide event rule patterns. Having reached maturity, the Big Data Systems and Internet of Things (IoT) technology require the implementation of advanced machine learning approaches for automation in the CEP domain. The goal of this research is proposing a machine learning model to replace the manual identification of rule patterns. After a pre-processing stage (dealing with missing values, data outliers, etc.), various rule-based machine learning approaches were applied to detect complex events. Promising results with high preciseness were obtained. A comparative analysis of the performance of classifiers is discussed.

Peer-review under responsibility of scientific committee of Missouri University of Science and Technology

Keywords: Complex Event Processing; Rule Based Classification; Machine Learning; Event Patterns

1. Introduction

Recent advancements in the Internet of Things (IoT), Big Data, and Cloud Computing technologies enable storing, processing, and analyzing continuously streaming low-level semantic data. Such data gathered from RFID tags and sensor devices, such as wireless sensors networks, pedometers, GPS, accelerometers, etc., offer huge

* Corresponding author. Tel.: +1-573-341-4749
E-mail address: enke@mst.edu

Peer-review under responsibility of scientific committee of Missouri University of Science and Technology
doi:10.1016/j.procs.2015.09.168

application opportunities. Literature analysis reveals the implementation of RFID and sensor data in diverse domains ranging from healthcare services, industrial production, military, and environmental applications, to social sensing that is based on observation of human behavior based on tracking information[1].

In spite of superiority in storage capabilities, computational complexity of streaming data prevents analysts from obtaining accurate results. The main issue related challenge to sensor data is the nature of its applications that require real-time analytics of huge uncertain data streams. Analysis of such continuous streaming was coined and investigated by researchers with various terms, such as "stream data processing", "information flow processing", "continuous data processing", and "high frequency data analysis", among others. Complex Event Processing (CEP) is one such emerging research domain that considers the streaming data as events and explores the possibilities to combine them for detecting unobservable actions.

CEP emerged as a novel technology with the purpose to identify complex events by analyzing, filtering, and matching semantically low-level events. The main idea behind CEP systems lies in real-time identification of situations by examining the cause/effect relationships among simple events that carry no specific information in stand-alone conditions. The result of such an analysis is pushing notification about detection of complex processes in real-time, which enable the decision makers to control the situation proactively and intervene flexibly in response to changing conditions. The effectiveness and usability of the CEP systems are measured in terms of the preciseness and timeliness of the decision-making.

Complex Event Processing has three core phases: (i) Filtering step: attempts to define the lists of the relevant simple events which can be attributed as predictors, (ii) Matching step: tries to identify the subset of simple events provided from the previous step by analyzing if they can fulfill the specific conditions of defined rule patterns, and (iii) Derivation step: detects more complex events using the information provided from matching subsets[2].

For current research, and for practical implementation, domain experts manually match event rule patterns for the detection of complex events. They are expected to choose the appropriate low-level events that are provided as inputs for rule derivation, define the relationship among the selected attributes, and provide the set of rule patterns for deriving more complex events. The complexity of deriving such rules depends significantly on the application domain and scenario. In some cases, domain experts easily provide such rules. Application of CEP in the financial sector is a typical example for these scenarios and can be presented as follows: "*Sell stocks of firms who have manufacturing facilities in USA and produce precious metals and have more than 5,000 employees and are at the moment in a reconstruction phase and their price/volume increase has been stable for the past 2 minutes*"[3].

In the presence of a vast amount of sensor and RFID data, it is unreasonable, and even not possible, to determine the matching rules manually by experts. Additionally, in many scenarios the rules can change and evolve dynamically depending on the specifications of the application domain[4]. Such uncontrollable changes make it not feasible for experts to constantly update their matching rules. Therefore, there is a need for automation of both processes: the initial identification of rules for detecting complex events, and updating them when the change in the behavior of the active systems is observed. The main research question of the underlying paper is the investigation of replacing rule derivation process from human judgments with machine learning approaches.

The remainder of the paper is structured as follows. In the next section we introduce the related literature analysis of current research direction in the CEP domain. After discussing the current situation, we provide the research question for the underlying approach. This section will be followed by a brief description of the rule induction based machine learning algorithm. After discussing the sensor data set, the empirical results of the model applications are presented. The paper will conclude with practical implications, shortcomings of the proposed work, and future research directions.

2. Related Work

Complex Event Processing is an emerging research domain that has attracted the attention of researchers in the last decade. A survey of the literature reveals that the researchers have mainly concentrated on the design and implementation of the novel systems that can handle complex event queries over streaming RFID and sensor data in real time. Jin et al.[5] implemented Timed Petri Nets to model their RFID complex event-processing engine. Researchers used Expressive Stream Language to combine the low level events and developed a special semantic for derivation of rule patterns for gathered primitive events. They discuss the effectiveness of their RFID CEP systems using real-world case studies. Wu et al.[4] proposed an automata-based event processing system that combines the filtering, combination, and correlation of events using SQL-like SASE language. After discussing the specifications of the proposed query language, the researchers illustrate their results with a case study from the retail and healthcare industry. Wang, et al.[6] developed a graph-based RFID complex event detection engine that they name RCEDA. The authors use temporal constraints in their model when detecting the complex events. They also mainly focus on the specification of the proposed query language. Wang, et al.[7] introduced a middleware architecture that integrates the wireless sensors networks and RFID systems and uses CEP in the proposed middleware to analyze the streaming data in real-time. The goal of the application is to provide reports for both customers and management through filtering, grouping, aggregating, and constructing complex events. The authors use Event Processing language to define the complex events. Liang and Dong[8] also developed a CEP system for managing and monitoring information systems through technologies such as Event Processing Language (EPL), caching strategy, and active database. They use CEP in their RFID middleware for identification of complex events. Cugola and Margara[9] designed a novel CEP system automata-based processing algorithm. In their application they use TESLA language to detect complex events from the primitive ones. The authors illustrate with a case study that their CEP system can process a large number of rules and offers a reduction of overhead costs. Terroso-Saenz and Valdes-Vela[10] created a CEP-based information middleware to develop real-time trajectory-based services. The pattern matching processes were executed through Esper technology and the highlights of the implementation using real and synthetic data were also introduced. Zang and Fan[11] proposed the architecture of event processing in enterprise information systems and event meta-modeling, defining the rules, operators, and other key components of the CEP system. They implemented the complex event processing mechanism in the enterprise information systems using RFID data and presented an architecture, optimization algorithm, and strategies.

3. Research Question

As can be inferred from the literature survey, scholars and practitioners have focused mainly on the design of querying complex event processes, application of various languages, and try to optimize the real-time analytics. All of these systems use the predefined rule patterns provided by experts. The lack of machine learning or statistical techniques in the current research, used to derive the rule patterns from the enormous size of streaming data, motivates us to investigate the integration of data mining approaches to CEP middleware.

Rule-based classifiers are special type of supervised learning which suit our purpose as the core idea is replacing manual definition of rule patterns. There are various types of rule-based classifiers having totally different algorithms and theoretical backgrounds. Therefore, there is a need to check the suitability of these methods when extracting rules from sensor and RFID data for CEP. The applicability of these techniques to be empirically validated in terms of diverse aspects, such as accuracy and reliability, among others, makes them worth investigating. This leads us to the primary question of our research:

Which rule-based classifiers are able to derive rule patterns from sensor data and what do the empirical results suggest about the performance of these models?

Since no prior research has adopted the rule-based classifier to automate the derivation of rule patterns in CEP systems, this research will discuss the integration of rule-based machine learning approaches to CEP systems. After providing a brief description of the selected models, the superiority of the classifiers will be verified using relevant statistical measures.

4. Rule-based Models

Various machine learning based classifiers can be implemented for determination of event patterns for streaming sensor or RFID data. As mentioned above, the main goal of this paper is induction of readable rules that can be easily interpreted by decision makers. Therefore, the rule-based classifiers were chosen in the underlying paper for identification of rule patterns to match events. Rule-based classifiers have already been successfully implemented in various information research domains such as human computer interaction[12], intrusion detection[13], content-based image retrieval[14], fingerprint identification[15], and machinery fault diagnosis[16], with each application domain delivering promising results. After providing a brief theoretical overview of the most widely used six rule-based classifiers, we will discuss the application these machine learning algorithms for identification of complex event processes from sensor data.

As proposed by Holte[17], One-R is one of the most widely applied rule-based classifiers due to its simplicity and agility. Based on a one level decision tree, this machine learning algorithm attempts to classify the instances by using the value of single attributes. In spite of the accepted lack in accuracy of this classifier, simplicity and speed of this approach make it as a crucial alternative to more complex rule based models[18]. Distinguished from other classifiers that use entropy measures to classify the instances, One-R classifier uses the error rate obtained from the training set. The proposed algorithm develops a rule for each individual predictor in the training set and determines the "one" rule with lowest error rate.

Also known as JRip, Repeated Incremental Pruning to Produce Error Reduction (RIPPER) was implemented by Cohen[19] to produce easily readable, fast, and accurate rules from noisy and large data sets. The main idea of RIPPER approach is for seeking an initial set of rules and iteratively improving it by applying an optimization algorithm. Such modelling with determination of initial rule sets makes this approach effective and fast. The training set used in the rule induction process of this approach is split into two parts: growing set and pruning set. The instances from the growing set are used to build a rule set that starts with an empty set. Once the rule is grown using the data from the first set, the instances from the pruning data set are applied to advance the performance of the obtained set by pruning it.

Frank and Witten[20] proposed an algorithm based on partial decision trees, PART, which differs from other alternatives in way that the rules are generated. The PART algorithm is a combination of C4.5 decision tree and RIPPER algorithms. Distinguished from other rule induction classifiers, the PART algorithm doesn't perform global optimization when inducing the rules, which makes it simple and fast. The working principle of this approach is based on separate and conquer strategy, as follows: the first rule is derived, instances covered by this rule are removed and recursively other rules are generated until there are no more instances remaining.

Proposed by Hall and Frank[21], DTNB is a combination of Decision Tables and Naïve Bayes approaches. The model is a Bayesian Network in which the conditional probabilities are represented with Decision Tables. In the DTNB algorithm, the attributes are divided into two subsets by applying the gain function. These two subsets are used to create Decision Tables and Naïve Bayes model, respectively. The algorithm is based on a forward selection procedure, where all attributes are initially modelled by Decision Trees and the selected attributes are provided to a Naïve Bayes model. In order to generate the overall class probability, the class probabilities estimated by Decision Tables and Naïve Bayes have to be combined.

Standing for Ripple-Down Rules, Ridor is a rule induction algorithm that is similar to PART and C4.5 approaches, but derives rules directly using Cendrowska's Prism algorithm in order to deal with noisy data. The Ridor approach is developed by Gaines[22]. The algorithm initially derives a rule that is followed by determination of exception to the defined rule using a least weighted error rate. For each exception the algorithm determines the most appropriate exception and this process continues recursively until all instances are covered. Derivation of exception can be also seen as tree algorithm where the exceptions are sets of rules for classification of classes.

Non-Nested Generalized Exemplars (NNGE) is an extension of Nested Generalized Exemplars (NGE), which is also an extension to the nearest neighbor classification approach that learns incrementally from the examples. NNGE was proposed by Martin[23] with the goal to solve the overgeneralization problem in the NGE method, which leads to the poor performance. The NNGE creates a new generalization each time a new instance is added to the system by distributing it to the nearest neighbor of the same class.

5. Data

In order to evaluate the ability of the selected machine learning approaches for inducing CEP rule patterns from large datasets, in this paper we use sensor data generated in phone-based accelerometers, which identifies the physical activities of users. The real-world dataset presented by the Department of Computer and Information Science of Fordham University, which was conducted by gathering accelerometer sensor data from 29 users and indicates their daily activities, was used for our empirical analysis[24].

Accelerometer sensors provide three important measures, namely (i) z-axis values, which record the forward movement of the leg, (ii) y-axis, which captures the upward and downward motion, and (iii) x-axis, which captures horizontal movement of the user's leg. Daily routine activities such as walking, jogging, ascending stairs, descending stairs, sitting, and standing were selected as outputs of the classification problem.

Before inducing rule patterns using the classifiers, the sensor data set has to be pre-processed. One of the most widely recognized and challenging issues deals with the missing values. There are diverse causes for missing values in the context of sensor data. The failure to measure the values due to abrasion effects and the problems related to the transfer of the data from sensor device to databases are two popular reasons for missing values. Kadlec[25] propose various approaches to deal with missing values in sensor data. Deleting the values of the missing instances, applying an algorithm based on iteratively reweighed least squares, and replacing missing values with the mean values of the variables are some typical approaches. In the underlying research, we adopt the later approach due to its simplicity and effectiveness.

Another important issue affecting the performance of the classifiers is data outliers. Outliers are sensor data that deviate from the relevant and logical values of the variables. These outliers can be obvious, which refer to the incorrect measurements of the sensor devices. Using the catalogue of meaningful ranges for variables, these outliers can be easily detected and removed. A challenging issue is identification of hidden outliers, which lay in the predetermined range but cannot be representative for the affected variable. In our research we implemented the "outliers" package of the R software and determined the list of instances with the most difference of the affected variable and removed them from the list.

6. Experimental Settings and Empirical Results

The main idea behind applying rule-based classifiers is detection of rule patterns in an offline mode and then providing the obtained rules to CEP engine, which can identify the complex process events from streaming data. As the behavior of streaming data changes over time, the rule pattern identification has to be conducted periodically in the offline regime and submitted to the CEP engine.

The accelerometer sensor data was collected every 50ms, which makes 20 measurements per second. A total of 80% of the 1,050,000 collected data points were used for testing each rule-based classifier, with the remaining 20% of the dataset used for testing purposes. In order to capture the temporal semantics in derivation of rule for recognition of complex processes, we extended the single time points of sensor values by including their lagged values. The values from the previous five periods for each three-sensor input were included to our dataset. By adding these attributes, the number of input variables reached 18.

After preprocessing data, a WEKA tool was used for classification of instances[26]. Various error measures, such as Root Mean Squared Error, Mean Absolute Error, Relative Absolute Error, and Root Squared Relative Absolute Error have been used to analyze and compare the performance of the selected rule-based classification approaches. Table 1 provides a summary of the error measures and classification accuracy. The highest accuracy belongs to the PART algorithm, which classified 93.14% of the instances correctly. Other methodologies, except One R, yield an average accuracy of 92%. The worst performance was shown by One R, which failed to classify slightly more than 20% of instances, which can be considered reasonable for application in detection of rule patterns for CEP systems.

Table 1. Error rates

Classifier Algorithm	Root Mean Squared Error	Mean Absolute Error	Relative Absolute Error	Root Relative Squared Error	Accuracy
One R	0.2847	0.0811	27.0347 %	73.5949 %	79.89%
RIPPER	0.1846	0.0614	20.3561 %	46.3709 %	92.13%
PART	0.1454	0.0383	12.7609 %	37.5896 %	93.14%
NNge	0.1823	0.0333	11.0306 %	46.9676 %	92.41%
Ridor	0.1796	0.0322	10.754 %	46.4164 %	91.87%
DTNB	0.1684	0.0533	17.7888 %	43.5178 %	89.72%

The analysis of error rates reveals that the ranking of the classifier algorithms is similar to accuracy raking, but there are minor differences in ranking according to individual measures. The PART algorithm show slightly better performance compared to the other models in error terms. RIPPER, NNge, and Ridor obtain close values that are slightly better than DTNB. The highest error rates can be observed in the results provided by One R approach. It is very important to figure out whether the classification results are reliable and were not simply obtained by chance or guesswork. Kappa's coefficient (Cohen's Kappa) is a statistical measure used for this purpose that is assumed to give the rating of the magnitude of agreement between observers. The value of Kappa statistics is measured as the difference between the observed agreement and expected agreement, which refers to occurrence by chance. The formula of the coefficient is as follows:

$$\mathrm{k} = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \tag{1}$$

where *Pr(a)* is the observed agreement and *Pr(e)* is expected agreement among raters. According to Landis and Koch[27], if Kappa's coefficient is equal to 0, there is chance, while if it is equal to 1, there is a perfect agreement among observers. The values between these extremes can be interpreted as follows: Slight agreement for 0.01 to 0.20, fair agreement 0.21 to 0.40, moderate agreement for 0.41 to 0.60, substantial agreement for 0.61 to 0.80, and almost perfect agreement for 0.81 to 0.99.

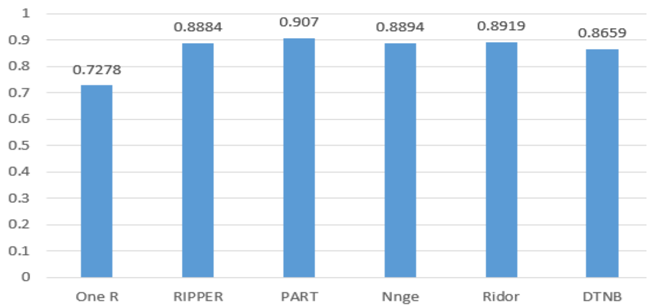

Fig. 1. Kappa Coefficients

Empirical results reveal that only the PART classifier exceeds the 0.90 threshold, which implies an almost perfect agreement among raters (See Figure 1). The PART classifier results are followed by the Ridor, NNge, RIPPER, and DTNB, with Kappa's coefficient of 0.8919, 0.8894, 0.8884, and 0.8659, respectively. These values suggest that the results provided by all these six algorithms are reliable. One-R indicates a medium level performance with 0.7278.

7. Conclusion

Detection of rule patterns from continuously streaming sensor and RFID data for matching primary events is a core task of CEP systems. Due to high velocity and the volume of these data, domain experts cannot provide the rules manually. Rule-based classifiers are the machine learning algorithms that can replace experts in generating rule patterns. In this underlying research we conducted an empirical study to investigate the applicability of rule induction algorithms in sensor data, and compared their performance using various error measures, classification accuracy, and Kappa values. High classification accuracy and low error rates suggest that rule-based classifiers can be used for detecting rule patterns in CEP systems. Kappa's coefficients validated that the results were not obtained by chance. The highest performance, in terms of classification accuracy, error rates, and Kappa values was obtained by the PART algorithm, which provides performance that is slightly better than the NNge, Ridor, and RIPPER algorithms. Application of fuzzy sets-based rule induction algorithms, in order to capture the uncertainties related to sensor data, is a potential future research direction. As mentioned above, these classifiers have to be trained in an offline mode. Application of novel algorithms that can directly handle streaming data in an online regime is also a subject for future research.

References

1. Aggarwal CC. An introduction to sensor data analytics. *Managing and Mining Sensor Data*. Springer, 2013, pp. 1-8.
2. Etzion O, Niblett P. *Event processing in action*. Manning Publications Co., 2010.
3. Teymourian K. Knowledge-based processing of complex stock market events. *Proceedings of the 15th International Conference on Extending Database Technology,* 2012. ACM.
4. Wu E, Yanlei D, Rizvi S. High-performance complex event processing over streams. *Proceedings of the International Conference on Management of Data,* 2006. ACM.
5. Jin X, Lee X, Kong N, Yan B. Efficient Complex Event Processing over RFID Data Stream. *Proceedings of the Seventh IEEE/ACIS International Conference on Computer and Information Science, 2008. ICIS 08.* pp.1-6, 14-16, May 2008.
6. Wang F, Liu S, Liu P, Bai Y. Bridging physical and virtual worlds: complex event processing for RFID data streams. *Advances in Database Technology-EDBT 2006*. Springer, 2006, pp 588-607.
7. Wang W, Jongwoo S, Daeyoung K. Complex Event Processing in EPC Sensor Network Middleware for Both RFID and WSN. *Proceedings of the 11th IEEE International Symposium on Object Oriented Real-Time Distributed Computing (ISORC), 2008*; pp. 165-169.
8. Liang D, Dong, W, Huanye S. Design of RFID Middleware Based on Complex Event Processing. *Proceedings of the IEEE Conference on Cybernetics and Intelligent Systems, 2006*; pp. 1-6.
9. Cugola G, Margara A. Complex event processing with T-REX. *Journal of Systems and Software* 2012; 85(8): 1709-1728.
10. Terroso-Saenz F, Valdes-Vela M, den Breejen E, Hanckmann P, Dekker R, Skarmeta-Gomez AF. CEP-traj: An event-based solution to process trajectory data. *Information Systems,* 2015; 52(0): pp. 34-54.
11. Zang C, Fan Y. Complex event processing in enterprise information systems based on RFID. *Enterprise Information Systems,* 2007; 1(1): pp. 3-23.
12. van den Bosch A, Krahmer E, Swerts M. Detecting problematic turns in human-machine interactions: Rule-induction versus memory-based learning approaches. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics,* 2001. Association for Computational Linguistics.
13. Wenke L, Stolfo SJ, Mok KW. A data mining framework for building intrusion detection models. *Proceedings of the IEEE Symposium on Security and Privacy,* 1999.
14. Jaworska T. Application of fuzzy rule-based classifier to CBIR in comparison with other classifiers. *Proceedings of the 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2014*; pp. 119-124.
15. Wu J, Shan JX, Dong-Hun S, Won DL. A new approach for classification of fingerprint image quality. *Proceedings of the 7th IEEE International Conference on Cognitive Informatics, 2008. ICCI 2008*, pp. 375-383.
16. Sun J, Xu X. Large Rotating Machinery Fault Diagnosis and Knowledge Rules Acquiring Based on Improved RIPPER. *Proceedings of Second International Conference on the Intelligent Computation Technology and Automation, 2009. ICICTA '09*, pp. 549-552.
17. Holte RC. Very simple classification rules perform well on most commonly used datasets. *Machine Learning,* 1993; 11(1): pp. 63-90.
18. Elomaa T. In defense of C4. 5: Notes on learning one-level decision trees. *ML-94* 2014; 254: 62.
19. Cohen WW. Fast effective rule induction. *Proceedings of the Proceedings of the Twelfth International Conference on Machine Learning,* 1995.
20. Frank E, Witten IH. Generating accurate rule sets without global optimization, 1998.
21. Hall M, Frank E. Combining Naive Bayes and Decision Tables. *Proceedings of the FLAIRS Conference,* 2008.
22. Gaines BR, Compton P. Induction of ripple-down rules applied to modeling large databases. *Journal Intelligent Information Systems,* 1995; 5(3): pp. 211-228.
23. Martin B. Instance-based learning: nearest neighbour with generalisation. University of Waikato Hamilton, New Zealand, 1995.
24. Kwapisz JR, Weiss GM, Moore SA. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter* 2011; 12(2): pp. 74-82.
25. Kadlec P, Gabrys B, Strandt S. Data-driven soft sensors in the process industry. *Computers & Chemical Engineering* 2009; 33(4): pp. 795-814.
26. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter, 2009; 11(1): pp.10-18.
27. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977, pp. 159-174.