CEP Rule Extraction from Unlabeled Data in IoT

Mehmet Ulvi Şimşek
Department of Computer Engineering Gazi University
Ankara, Turkey
mehmet.ulvi.simsek@gazi.edu.tr

Suat Özdemir
Department of Computer Engineering *Gazi University*
Ankara, Turkey
suatozdemir@gazi.edu.tr

***Abstract*—With the recent development of the Internet of Things, produced data are increasing day by day. These data have to be analyzed in real time. To provide real time analysis, Complex Event Processing is proposed to analyze the continuous and timely annotated data. Complex event processing detects complex events from atomic events via predefined rules which are mostly determined by domain experts. Determining complex event processing rules requires thorough knowledge of the data and data relations among data sources. It will be difficult to define a rule when it is considered that the scope and quantity of data is increased. Therefore, there is a need for extracting rules automatically. In this paper, we propose a novel model that extracts rules from unlabeled data by using clustering and rule mining algorithms. The model is evaluated in terms of classification performance and the results show that the proposed model is a promising solution for extracting complex event processing rules.**

Keywords—Complex Event Processing, rule extraction,

I. INTRODUCTION

Devices such as sensors, mobile phones etc. have emerged to explosion in the volume, velocity and variety of the data [1]. The vast amount of data are produced by the devices with the advent of Internet of Things (IoT) [2]. The data is continuous and timely annotated data is to be analyzed in real time. Complex Event Processing (CEP) emerged as promised technology with achieving situational awareness and detecting events in real time [2].

CEP analyzes large flows of timely annotated stream data received from a monitored environment to detect complex events in real time. It can be employed in different domains such as social networks, traffic monitoring, crisis management etc [3,4]. In more detail, CEP is a service that matches atomic events to complex events [5]. The events are thought as composite of atomic events (complex event) in time. CEP systems differ in many aspects such as architecture, data models, rule languages, and processing mechanisms [2]. The system detects complex events via predefined patterns and rules [6]. The incoming stream data is analyzed by CEP rules to find which is relevant or not. Any data that does not match the rules is ignored.

CEP rules constitute the main part of the system. These rules which are mostly determined manually by domain experts manage CEP engine. The rules are written manually by experts that are required to have thorough knowledge of the event. Therefore, it is hard to define rules when the event type in IOT will increase. The rules can also change in the application domain. Uncontrollable changes require experts to update the rules detecting the complex events. Therefore, instead of manually defining the rules, they need to be extracted automatically from event history. This paper proposes a novel model that extract rules from unlabeled data in CEP domain. In our model, IoT data is clustered and then, analyzed to extract rules using rule mining algorithms from event history.

The rest of the paper was structured as follows. We first introduced the related works in Section II. Then, we gave essential information about CEP rules in Section III. We presented the extraction model of CEP rules and the evaluations of the experiment in Section IV. Finally, we concluded the paper in Section V.

II. RELATED WORKS

Although many research have been made in the area of CEP, rule mining in this domain have not been studied as much as processing and distributed processing. Many researchers have concentrated on processing performance. Due to the advance of the IoT, their applications generate vast amount of data by monitoring the physical phenomenon surrounding them. The generated data are heterogeneous, distributed and continuously changing and increasing day by day; hence, maximizing the data analysis efficiency in such data is important [7]. In this context, automated rule extraction is important for non-experts to deploy rules automatically.

CEP rules have many parameters such as window size, sequence, operator etc. For this reason, determining the window size and other issues are important to extract CEP rules. In [4,8], they separated five sub problems to extract rules as follows. i) Window learner, ii) Events and attributes learner, iii) Predicates learner, iv) Sequence learner, v) Negations learner. Similarly, Frömmgen [9] proposes genetic algorithm model to learn the rules for adaptive distributed systems. The system generates rule sets with modular architecture and strategies such as correlation of events, monitoring values, adaptations actions and resulting utilities. In [10], the authors used a noise Hidden Markov Model (nHMM) to infer rules automatically. The event tagged by domain experts and the model learn the pattern detector.

In [11], the authors aimed to attain a machine learning based approach for extracting rules or classifying from historical data. After the preprocessing stage for extracting feature, rule based algorithms have been used to detect complex events. PART algorithm shows the highest accuracy when compared with others such as ONE-R, RIPPER, DTNB, Ridor, NNge. It shows that rule based algorithms can be used as generating rule patterns with high accuracy. In [3], the authors used shapelets learning to extract pattern of minimum possible length. After that, the extracted shapelets

978-1-5386-7893-0/18/$31.00 ©2018 IEEE

were used for transforming to CEP rules. Machine learning based approach has a great performance in extracting rules, but in [12], it was stated that fuzzy-based systems could give better results when considering sensor problems such as conflicting and missed readings. After selecting feature subset, Fuzzy Unordered Rule Induction Algorithm (FURIA) has been used for identifying event patterns.

III. CEP RULES

CEP is used to extract meaningful information from multiple stream of data in real time [13]. Its architecture consist of event observers (sources), event consumer (sink), CEP engine and rule managers (Fig. 1). The engine is the main part of the architecture that operate rules from conceived. In addition, CEP engine has its own data model and rule definition language and processing algorithm and implementation [14].

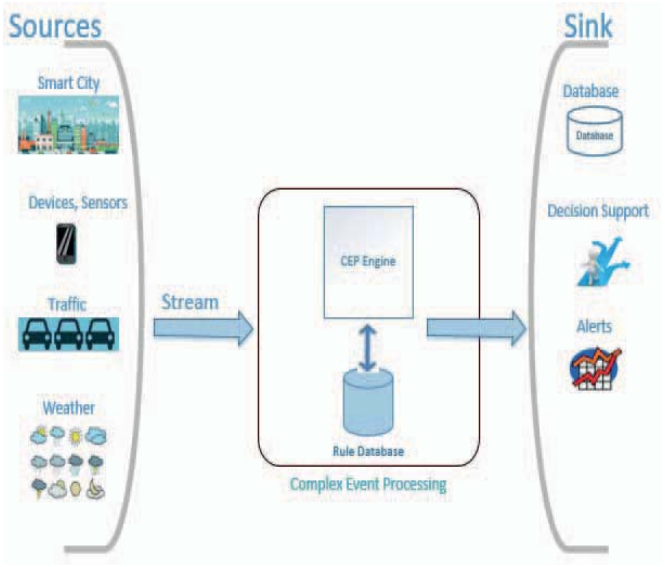

Fig. 1. CEP architecture

Rule languages named as event query languages grouped into three categories according to styles as Composition Operators, Data Stream Query Language and Production Rules. In more detail, different composition operators such as conjunction, sequence, negation are used in composition operators. Data stream query language used like SQL queries that are evaluated according to event relations. Production rules execute the actions based on certain condition job [15]. The rule languages consist of operators. CEP operators like windowing, selection, sequence are CEP rule syntax that forms CEP rules [17]. An example for such SQL like language is as follows [16]:

SELECT * FROM (t S1 ; t S2 ; t L) WHERE FILTER(S1.acc = S2.acc), FILTER(S2.acc = L.acc), S1.amount < 100 AND S2.amount < 100 AND L.amount > 250 AND (S1, L) OCCURS WITHIN 12 hours

There are several languages in which their concepts, mechanism and structure are the same [18]. Generally, these rules are defined by experts. The experts define the rules to detect complex events to match the rule pattern. Therefore, they are required to have a knowledge in application domain and scenario. Authors define the rule extracting methods as follows [19]:

- Domain experts
- Rule mining
- Mathematical optimizations

When the complexity is manageable, domain experts can define rules. If the data contains relevant information and stored in database, you can use rule mining algorithm. In the last approach, mathematical optimizations were used in a complexity going beyond human capabilities [19].

Heterogeneous data collected from sensor devices are increasing day by day. In the future, experts will not recognize the relations of the data with each other. Therefore, manually writing rules is a challenging task for humans [11].

IV. PROPOSED RULE EXTRACTION MODEL

In this paper, a model was proposed to extract rule information from historical data. The proposed model ensures rule extraction automatically without any support of experts.

In this model, stream of data comes from different sources such as smart city, devices, sensors, traffic and weather sensors. The data has been stored in the historical database and analyzed in the CEP architecture. In addition, CEP analyzes the data stream in near real time with rules. The rules determine which pattern to detect or not in the CEP architecture. The output of the CEP architecture feeds other systems such as database, decision and alarm systems.

The proposed model consists of two stages and the complete structure of the proposed model was shown in Fig.2.

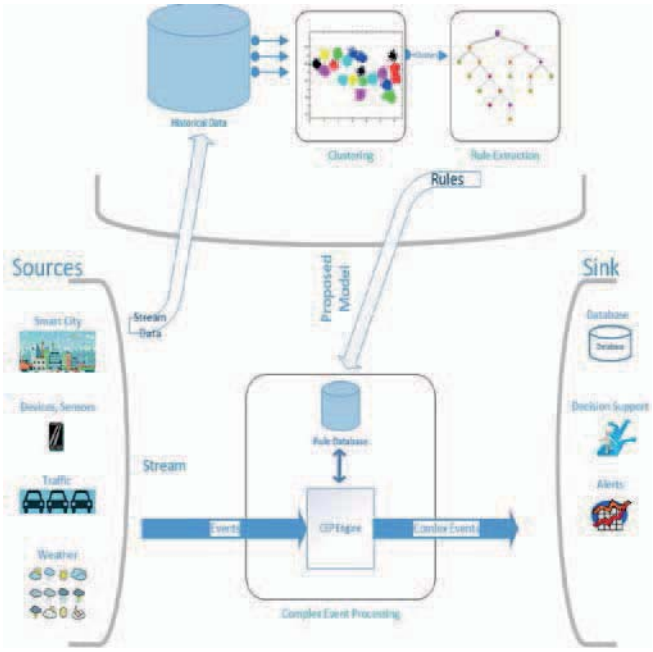

Fig. 2. The proposed architecture

CEP architecture ignores the data that does not match the rules. However, all data are stored in the historical database

for analysis. The data does not consist a class attribute. To extract rule information, it is important to see the relationships of data with each other. In the first stage, we used a cluster algorithm for labeling the K-means algorithm was used in this stage to partition into k clusters.

In the second stage, the clustered data was performed by rule extraction algorithms to extract rules. Training was carried out by PART, JRIP, Decision Table and ONE-R algorithms for the clustered data. After extracting the rules, the extracted rules were stored to the rule database for using in CEP engine.

A. Dataset

CEP is used in different domains especially in IoT data coming from sensors. In this study, we used a pollution dataset which is collected for City Pulse EU FP7 Project [20]. The project provides a smart city application in IoT. It consists of road traffic, pollution, weather, cultural, library and parking data that were collected from the cities of Aarhus and Brasov in Denmark and Romania, respectively between 2013 and 2015.

In this model, we used a pollution dataset for the experiment. The pollution dataset contains 5 features including ozone, particulate matter, carbon monoxide, sulfur dioxide, nitrogen dioxide. The dataset has 17568 samples that were collected at five minute intervals. Each sample value is given in type of air quality index standard and ranges from 1 to 215. One of the examples of these gases stream was shown in Fig. 3.

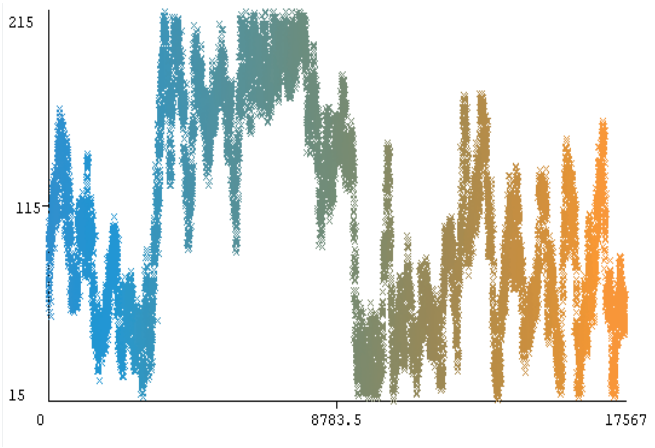

Fig. 3. Nitrogen dioxide data frame

B. Clustering And Rule Classifiers

Clustering is used for finding groups or clusters of similar objects in multivariate data. Each group that are called cluster has similar characteristics within itself. In clustering analysis, having a knowledge about the number of clusters defines the model accuracy. Hence, the cluster size identification is an important task that can be determined by several approaches such as Elbow method, Information Criterion Approach, An Information Theoretic Approach, Choosing k Using the Silhouette and Cross-validation [15].

A Canopy clustering algorithm is a pre-clustering algorithm which is performed before a K-means clustering. The Canopy algorithm is used for defining the initial clustering center and optimal value k of the data which is used as input parameters of the K-means algorithm [21]. After Canopy cluster algorithm was realized, optimal k value obtained to cluster the data for performing rule extraction algorithms.

After the k value was determined, K-means algorithm was used to label data. K-means clustering has been used in many areas to partition a data set into k groups. It starts with the center of the cluster and firstly each sample placed arbitrary position. Then, changes the center to minimize the clustering error [22].

Euclidean distance was selected for measuring the distances which determine the objects' cluster labels [23]. Parameters of the network were denoted in Table I.

TABLE I. PARAMTERS OF K-MEANS CLUSTER

Parameters	Values
canopyMaxNumCanopiesToHoldIn memory	100
canopyMinimumCanopyDensity	2
Distance Function	Euclidean Distance
Max Iteration	500
Number of Cluster	8

Data mining have been used to extract information and patterns for decisions making [24]. Rule based classification is a branch of supervised classification in data mining. Rule based classifiers are used for identifying the rule patterns to match events [11]. Rule extraction algorithms such as PART, JRip, Decision table and ONE-R has been carried out in training that extracts rules with the highest accuracy rate.

C. Performance Evaluation

In this section, we introduced the results of evaluation of rule extraction model. First, to obtain the optimal k value, the Canopy algorithm was trained and Canopy clustering results were shown in Table II.

TABLE II. CANOPY CLUSTERING RESULT

Algorithms	
Canopy Clustering	Number of canopies (cluster centers) found: 8 T2 radius: 0,818 T1 radius: 1,023

The k value was obtained as 8 and used as input parameters for K-Means clustering. We evaluated the clustering performance with sum of squared errors as shown in Table III.

TABLE III. K-MEANS CLUSTERING RESULT

Algorithms	k	Within cluster sum of squared errors
K-Means	8	2292.75

Second, we compared the rule extraction algorithms in the model to verify the prediction ability. To evaluate the prediction performance, we used Precision, Recall and F1-Score metrics. We evaluated the results in this context, the best Precision (%97,8), Recall (%97,8) and F1-Score (%97.8) results were achieved on PART and JRIP algorithms

with our model. Compared the other two algorithms, the decision table presented underperformance. ONE-R algorithm showed the worst performance when compared with other algorithms.

TABLE IV. RULE BASED ALGORITHMS RESULT

Algorithms	Precision	Recall	F- Measure
PART	0,978	0,978	0,978
JRip	0,978	0,978	0,978
Decision Table	0,832	0,816	0,820
ONE-R	0,343	0,402	0,35

On the other hand, we evaluated the rules in terms of extracting number, it was seen that decision table produced more rules than the other algorithms. The extracted number of the rules was shown in Table V.

TABLE V. EXTRACTED NUMBER OF RULES

Algorithms	Number of Rules
PART	102
JRip	76
Decision Table	2771

Extracted rules defines the boundaries of the cluster and similar to SQL language. The extracted rules are interpreted easily by domain experts. One of the examples was shown as follows:

```
Sulfure_dioxide <= 56 AND
particullate_matter <= 133 AND
nitrogen_dioxide <= 142 AND
nitrogen_dioxide > 70 AND
carbon_monoxide <= 108: cluster0

ozone > 130 AND
nitrogen_dioxide > 167 AND
carbon_monoxide > 54 AND
nitrogen_dioxide <= 196 AND
sulfur dioxide <= 127 AND
carbon_monoxide <= 130: cluster4
```

CONCLUSIONS

IoT produces a vast amount of data which can be new heterogeneous data. These data are particularly collected from sensor devices. In CEP domain, writing rules for this type of IoT data is a challenging task. Therefore, since extracting rule patterns from IoT data is challenging, the rules need to be extracted automatically.

In this paper, we proposed a novel model to address the CEP rule extraction in unlabeled IoT data. First, Canopy algorithm was used to define optimal k value and the data was clustered by K-Means algorithm. Second, rule mining algorithms were used for the identification of rule patterns to match events. They were evaluated in terms of Precision, Recall, F1-measure. In the light of the findings, the results show that the rule extraction model from unlabeled data is promising. The model can easily extract rules instead of domain expert-defined rule. In future, we will search deep learning algorithms to extract hidden patterns in CEP domain.

REFERENCES

[1] M. D de Assuncao, A. da Silva Veith, R. Buyya, "Distributed data stream processing and edge computing: A survey on resource elasticity and future directions," Journal of Network and Computer Applications, 2018,103, 1-17.

[2] F. Starks, V. Goebel, S. Kristiansen, T. Plagemann, "Mobile Distributed Complex Event Processing—Ubi Sumus? Quo Vadimus?," In Mobile Big Data (pp. 147-180). Springer, Cham, 2018.

[3] R. Mousheimish, Y. Taher, K. Zeitouni, (2016, June). "Automatic learning of predictive rules for complex event processing: Doctoral symposium," In Proceedings of the 10th ACM International Conference on Distributed and Event-based Systems (pp. 414-417),ACM, June 2016.

[4] A. Margara, G. Cugola, and G. Tamburrelli. Learning from the past: automated rule generation for complex event processing. In Proceedings of the 8th ACM International Conference on Distributed Event-Based Systems, pages 47–58. ACM, 2014.

[5] D. Robins, "Complex event processing," In Second International Workshop on Education Technology and Computer Science. Wuhan (pp. 1-10), February 2010.

[6] I. Flouris, N. Giatrakos, A. Deligiannakis, M. Garofalakis, M. Kamp, M. Mock, "Issues in complex event processing: Status and prospects in the big data era," Journal of Systems and Software, 127, 217-236, 2017.

[7] M. I. Ali et al., "Real-time data analytics and event detection for IoT-enabled communication systems," Web Semantics: Science, Services and Agents on the World Wide Web, vol. 42, pp. 19-37, 1// 2017.

[8] A. Margara, G. Cugola, and G. Tamburrelli. Towards automated rule learning for complex event processing. Technical report, Technical Report, 2013.

[9] A. Frömmgen, R. Rehner, M. Lehn, A. Buchmann, "Fossa: Learning eca rules for adaptive distributed systems," In Autonomic Computing (ICAC), 2015 IEEE International Conference on (pp. 207-210). IEEE, July 2015

[10] C. Mutschler and M. Philippsen, "Learning event detection rules with noise hidden markov models," In Adaptive Hardware and Systems (AHS), 2012 NASA/ESA Conference on, pages 159–166. IEEE, 2012.

[11] N. Mehdiyev, J. Krumeich, D. Enke, D. Werth, P. Loos, P. "Determination of rule patterns in complex event processing using machine learning techniques" Procedia Computer Science, 61, 395-401, 2015.

[12] N. Mehdiyev, J. Krumeich, D. Werth, P. Loos, "Determination of event patterns for complex event processing using fuzzy unordered rule induction algorithm with multi-objective evolutionary feature subset selection. In System Sciences (HICSS), 2016 49th Hawaii International Conference on (pp. 1719-1728). IEEE, January 2016.

[13] A. S. Al-Haboobi, A. F. Alharan, R. H. Alsagheer, "Experimenting with Storm/Esper Integration and Programmatic Generation of Storm Topologies," International Journal of Computer Science and Information Security, 14(8), 169, 2016.

[14] G. Cugola, A. Margara, M. Matteucci, G. Tamburrelli, "Introducing uncertainty in complex event processing: model, implementation, and validation," Computing, 97(2), 103-144, 2015.

[15] T. M. Kodinariya, P. R. Makwana, " Review on determining number of Cluster in K-Means Clustering," International Journal, 1(6), 90-95, 2013.

[16] N. P. Schultz-Møller, M. Migliavacca, P. Pietzuch, "Distributed complex event processing with query rewriting," In Proceedings of the Third ACM International Conference on Distributed Event-Based Systems (p. 4), July 2009.

[17] R. Mousheimish, Y. Taher, K. Zeitouni, K. " Automatic learning of predictive cep rules: bridging the gap between data mining and complex event processing," In Proceedings of the 11th ACM International Conference on Distributed and Event-based Systems (pp. 158-169), June 2017.

[18] L. Burgueño, J. Boubeta-Puig, A. Vallecillo, " Formalizing Complex Event Processing Systems in Maude" IEEE Access. Pielmeier, J., Braunreuther, S., & Reinhart, G. (2018).

[19] J. Pielmeier, S. Braunreuther, G. Reinhart, " Approach for Defining Rules in the Context of Complex Event Processing," Procedia CIRP, 67, 8-12, 2018.

[20] T. C. Consortium, "CityPulse Annual Report," The CityPulse Consortium, 2016.

[21] G. Zhang, C. Zhang, H. Zhang, " Improved K-means algorithm based on density Canopy" Knowledge-Based Systems, 145, 289-297, 2018.

[22] A. Likas, N. Vlassis, J.J. Verbeek, "The global k-means clustering algorithm," Pattern recognition, 36(2), 451-461, 2013.

[23] W. Zhang, H. Hu, H. Hu, J. Fang, " Semantic distance between vague concepts in a framework of modeling with words," Soft Computing, 1-18, 2018.

[24] A. Mahajan, A. Ganpati, "Performance evaluation of rule based classification algorithms, " International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Vol, 3, 3546-3550,2014.