IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 1, July 2023

Pattern Based Sequence Classification

Ms. Reshma Hasan Atar[1] and Dr. D. S. Bhosale[2]

Student, ME (Computer Science Engg.)[1]

Associate Professor, Computer Science Engg.[2]

Ashokrao Mane Group of Institutions, Vathar Tarf Vadgaon, Kolhapur, India

Abstract: *The text highlights the challenges of analyzing massive amounts of data, particularly in the context of business processes using process mining. Outliers or irregular behavior in the data can negatively impact processing and clutter process models, leading to less useful paths. The objective is to automatically extract process models from the data, and an automated method for removing irregular behavior from event logs is introduced. This method significantly improves the quality of identified process models and scales well to large datasets. Since the effectiveness of filtering strategies depends on the event log, users can interactively filter activities and directly view the filtered process model from the event log using a slider-based approach. Ultimately, the choice of included activities is left to the user. The method is tested using actual occurrence log collections from enterprise process oversight and hospital environments. The results demonstrate that the newly proposed activity filtering approaches yield process models that are more behaviorally specific compared to conventional frequency-based filtering methods***.**

Keywords: Bigdata, Process mining, Outliers, Process model, Event logs, Filtering strategies, Frequency-based filtering

I. INTRODUCTION

A sequence is an ordered list of single tones, and it can be found in a variety of sig- nificant contexts, including text, films, audio signals, biological structures, and web activity logs. The act of classifying new sequences is known as sequence classification. For sequence categorization, the knowledge acquired during the training phase is applied. To create classification rules, intriguing patterns a dataset of sequences that have labels are exploited. A sequencing issue categorization is solved adopting this regulation. There are two basic categories that apply to the majority of datasets utilised for the sequence classification challenge. In the first scenario, certain elements that co-occur inside a sequence though not always in the same order determine its class. In contrast, the items that make up a sequence determine the class of that sequence always appear similarly ordered the sequence. In indicated situation, It is ineffective to use classifiers that depend on sequential patterns since The right rules won't be found. Sequence-based classifiers ought to perform better than itemset-based classifiers in the first scenario. When the pattern occasionally occurs in a sequence that deviates from the norms, itemset-based classifiers will perform better. The aforementioned finding serves to motivate the proposed system, Pattern Based Sequence Classification, which includes various phases. In the first step, an approach is presented to detect both itemset and sub sequence interesting patterns. The second phase is the conversion of the patterns found into classifi- cation rules. In third step, offer two approaches for developing classifiers that can determine which category the latest example belongs torules in the first method are selected on the basis of confidence. In next method, it measure how close the items to with the dimensions of the smallest periods with patterns in various sequences, they are averaged against the previous one. Finally Evaluate the quality of patterns as features by using feature based classifiers. In addition, the system replaces classifiers based on Classification Based on Association technique (CBA) using an innovative classification that is depending on the HARMONY the score behave delivers higher performance. Algorithms are used to detect two different types of intriguing patterns, itemset and sub sequences. Additionally, it employs top-k method for every classifier supplied rather than just the highest ranking rule.

The system offers a fresh approach to feature vector representation that turns each sequence into a feature vector. The performance of the novel feature vector representation approach beats that of the previously published ones when using algorithms for sequence classification. Obtaining helpful process information from event logs is the aim of process mining. The wider field of process analysis includes several areas of interest, including process research, and these

DOI: 10.48175/IJARSCT-12057

deals with obtaining model processes from event logs. The complexity was generated procedure system and the degree to which these range algorithms effectively reflect the behaviour shown in a log are various trade-offs. The underlying premise of algorithms for process discovery is that an event log accurately captures a business method' behaviour as they were carried out in an entity over a specific time duration. regrettably, outliers frequently appear in real-life process event logs as well as other types of event logs.

The inability to effectively identify and filter out rare behaviour has a detrimental effect on the quality of the discovered model, particularly its accuracy, which is an assessment of the accuracy with which the model captures actions that weren't recorded in the log. In order to solve the issue of finding high-quality process models in the presence of noise, this work contributes an automated method for systematically removing irregular behaviour from event logs. The process behaviour that was recorded in the log is abstracted by our filtering method into an automaton (a directed graph). This automaton records the straight- line relationships between event labels in the log. The automaton is then stripped of infrequent transitions. Then, to find occurrences that no longer match, the original log is replayed on this condensed automaton. These instances are removed from the log. The technique aims to eliminate as many infrequent transitions from the automaton while removing as few events from the log as possible. The result is that the filtered log exactly matches the automata. The technique has been developed on top of the Framework and rigorously tested in conjunction with various baseline discovery techniques. It uses a three-pronged strategy.

First, to begin with, we tested the effectiveness of our method by injecting it into simulated logs at various noise levels. Evaluated the improvement of discovery accuracy and

Second, evaluated the improvement of discovery accuracy and reduction of process model complexity in the presence of varying degrees of noise by comparing the results of a variety of baseline process discovery algorithms with those obtained by two baseline automated filtering approaches.

Third, conduct this latter experiment once more using a variety of real-world logs with varying size and number of (separate) occurrences. Different structural complexity measures, such as size, density, and control- ow complexity, were employed as stand- ins for model complexity, with discovery accuracy quantified in terms of the well- established measures of fitness and precision

II. LITERATURE REVIEW

B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," in Proc. ACM SIGKDD Int. Conf. Knowledge. Discovery Data Mining, 1998, pp. 80–86.

The authors of this research suggested a framework for combining based on rules categorization and association mining. Several issues with the current classification system can be resolved with integration. To create all of the class association rules needed to create an accurate classifier, an algorithm is described. This method produces classifiers that are more precise than the most recent classification scheme.

Z. Xing, J. Pei, and E. Keogh, "A brief survey on sequence classification, "ACM SIGKDD Explorations Newsletter., vol. 12, no. 1, pp. 40–48, 2010.

The author provides a quick overview of the prior research on sequence categorize- on in this publication. The three types of sequence classification methods in this approach are feature-based, sequence distance-based, and model-based. Additionally, the author offers various extensions to the traditional sequence classification. Finally, the author compares all categorization techniques used in various application fields.

C. Zhou, B. Cule, and B. Goethals, "Item set based sequence classification, "in Ma- chine Learning and Knowledge Discovery in Databases. New York, NY, USA: Springer, 2013, pp. 353–368.

In this study, confident classification rules are produced using discovered item sets. Gives two additional methods for creating a classifier. The CBA approach is the foundation of the first classifier. The classification algorithm evaluates rules' importance in relation to towards the latest information protest ranking them. This Consequently, the approach used for identifying a series of data is effective and reliable

N. Lesh, M. J. Zaki, and M. Ogihara, "Scalable feature mining for sequential data," IEEE Intell. Syst., vol. 15, no. 2, pp. 48–56, Mar./Apr. 2000

This system's approach is to look through the pool of potential features and mine for those that are regular, predictable, and non-redundant. The sequence dataset effectively selects features. This method constructs a classifier using frequent and reliable patterns.

B. Cule, B. Goethals, and C. Robardet, "A new constraint for mining sets in sequences," in Proc. SIAM Int. Conf. Data Mining, 2009, pp. 317

By combining the stability as well as the frequency of its outline with an interestingness metric, this method creates a new restriction. There are provided algorithms for quickly finding specified interesting patterns.

J. Wang and G. Karypis, "Harmony: Efficiently mining the best rules for classification," in Proc. SIAM Int. Conf. Data Mining, 2005, pp. 205–216.

This method discovers employing the most recent collection of standards for classification a new classifier called Harmony. To select the rules with the best degree of confidence, an instance-centric rule generation method is utilized. These rules are then incorporated into the final rule set, improving the classifier's accuracy.

V. S. Tseng and C.-H. Lee, "Effective temporal data classification by integrating sequential pattern mining and probabilistic induction," Expert Syst. Appl., vol. 36, no. 5, pp. 9524–9532, 2009.

In this study, the author suggests the classify-by-sequence method, a pattern-based mining technique, for categorizing big temporal datasets. Sequential pattern mining and probabilistic induction are combined using CBS. simulators are made to assess CBS performance. CBS efficiently classifies the data as a result.

III. PROBLEM STATEMENT

Designing a system for sequence classification using interesting patterns by combing support and cohesion methods which gives better performance and also provides higher classification accuracy compared to other methods. Classifiers which build in this system uses HARMONY scoring function and also uses top-k strategy instead of using only highest ranked rules. A new feature vector representation approach is used to transform each sequences into a new feature vector.

3.1 OBJECTIVE

The Objectives of proposed work are as follows:

- To develop the pattern-based sequence classifier for improving the accuracy in interesting pattern mining.
- To develop the technique for measuring the interestingness of the pattern from the sequence in dataset.
- To improve the sequence classification by using methods cohesion and support of the pattern
- To improve the sequence classification for the events in the sequence that occurs at the same timestamp.

IV. METHODOLOGY

4.1 System Architecture

The performance of business processes is typically documented within the application record as well as an event record for the purposes for analysis and auditing. A A set of evidence are discovered in an event record. Each trace is nothing more than a process instance's log footprint. Sequence of occurrences is identified by a special case identifier. An event is used to describe how a certain process task is executed within a trace.

4.2 Detection Approach

We offer a method. The detection of irregular behaviour depends on the log automaton technique's detection of abnormalities. Anomalies reveal relationships between rare occurrences. An automaton is nothing more than a directed graph, where each state represents a job that might occur in a process event log and each arc between two states is a direct follow dependency between respective tasks.

DOI: 10.48175/IJARSCT-12057

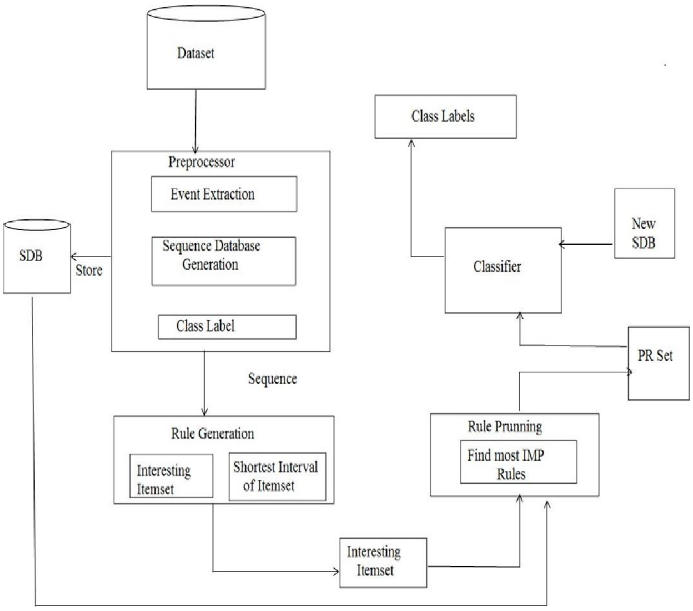
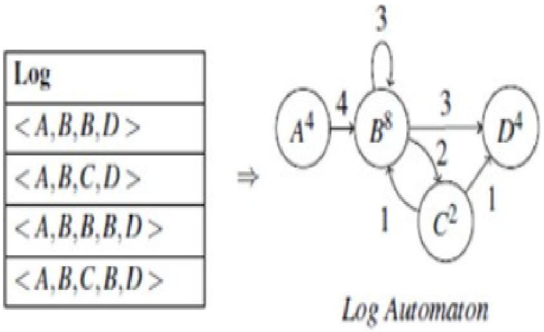


Fig 1. System Architecture

4.3 Infrequent Behavior Detection

The process event log is used to create a log automaton. Each log event is transformed. into states, along A initial and D final states. Additionally, a couple of states connect once via an arc these both record Observe one another. Ultimately, every phase and arc received a marking expressing the frequency.



Log
<A,B,B,D>
<A,B,C,D>
<A,B,B,B,D>
<A,B,C,B,D>

Figure 2: Example of Log Automaton REMOVAL OF

4.4 Uncommon Behavior

The process event log's sporadic behaviour results in events being logged in the wrong sequence and at the wrong time. Such irregular behaviour results in direct follow dependencies that are derived but never store or result in lead after relationships. Therefore, eliminating irregular behaviour means concentrating on identifying irregular arcs or inaccurately recorded occurrences.

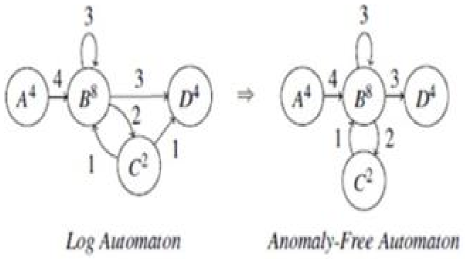


Figure 3: Example of Anomaly free Log Automaton

DOI: 10.48175/IJARSCT-12057

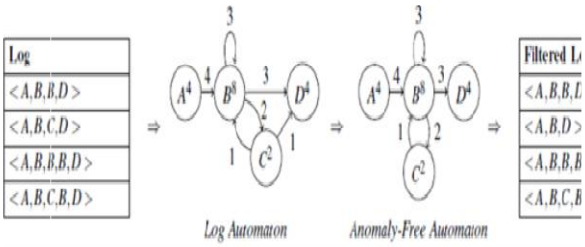


Figure 4: Anomaly free automaton to generate a log.

The following record in the filtrated an event C is removed. As an automated system with uncommon behavior and no deviations, it has reached its final stage of existence could not regenerate C event. Figure 3: Example of Anomaly free Automaton with Filtered Log Identification of Frequency Threshold To determine the ideal frequency threshold, lower-half interval, and upper-half interval ideas must be rendered. The interquartile range (IQR), which is nothing more than the difference between the upper and lower quartiles, is what determines IQRL and IQRU. Q1 and Q3 are the first and third quartiles, respectively. The difference between the median and lower quartiles (IQRL = median - Q1) is what makes up the lower half interquartile range. In a similar manner, the upper-half interquartile range (IQRU) is calculated as the difference between the third quartile and the median. An esti- mation of the arc frequency distribution curve's skewness is provided by the ratio between these two ideas, where rIQR = (IQRU)=IQRL > 1 indicates a positively skewed distribution. The frequency threshold that produces an arc frequency distribution curve with rIQR1 is the best in this case because it eliminates the fewest infrequent arcs. The frequencies of the remaining arcs change when infrequent arcs and the resulting events are eliminated, which changes the arc frequency distribution curve. Therefore, until no more events are removed, we suggest repeating the log filtering several times using the filtered log as input.

V. TOOL

The open-source platform for process mining methods is identified as ProM, which sig- nifies Process Mining Framework. A free tool that supports the creation of procedure mining extensions is known as ProM 6. There are numerous process mining extensions in this tool..

Datasets BPI Challenge 2012.

Events pertaining to an application procedure for an individual loan or an overdraft inside a Dutch banking system are tracked into the BPI Demand 2012 event log. There are 262, two hundred incidents, along with 13,087 cases in the event log. The application is evaluated financially after which it is either accepted and triggered, rejected, or canceled.

VI. RESULT AND DISCUSSION

The graph you are referring to shows the characteristics of the logs used in an experiment. The information includes the number of traces, number of events, number of unique labels, and the percentage of infrequent behavior.

Number of traces: A trace is a sequence of events that represent a single execution of a process. The number of traces in the logs ranges from 617 to 46,616. This means that the logs contain traces of different lengths, from very short to very long.

Number of events: An event is a single unit of information in a log. The number of events in the logs ranges from 9,575 to 422,563. This means that the logs contain a large number of events, from a few thousand to over a million.

Number of unique labels: A label is a category that is assigned to an event. The number of unique labels in the logs ranges from 9 to 22. This means that the logs contain events that are categorized into a variety of different groups.

Percentage of infrequent behavior: Infrequent behavior is behavior that occurs rarely in the logs. The percentage of infrequent behavior in the real-life logs varies from 2% to 39%. This means that in some of the logs, a significant proportion of the events are infrequent.

The wide range of characteristics in the logs shows that they are a diverse set of data. This diversity makes them a good challenge for log analysis algorithms. The algorithms need to be able to handle logs of different lengths, with different numbers of events and labels, and with different levels of infrequent behavior.

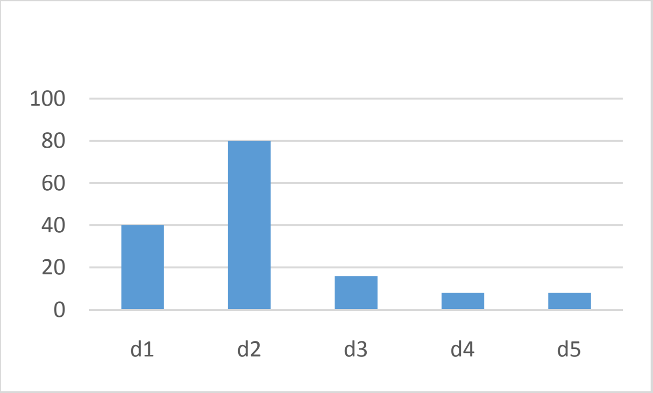


Figure 5: Infrequent Behaviour In Percentage

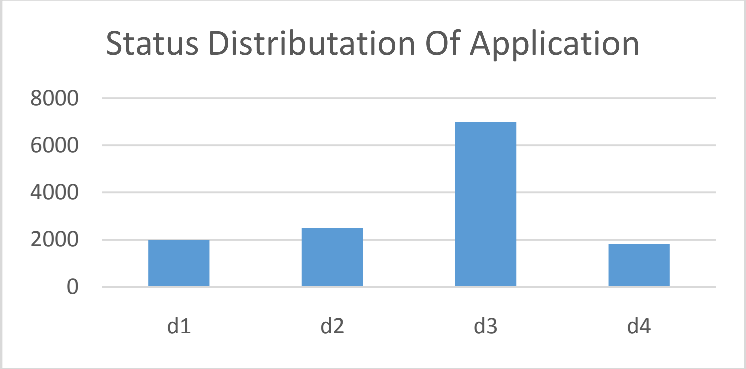


Figure 6: Status Distribution

VI. CONCLUSION

Presented a technique for removing irregular behavior from process execution records auto matically. The fundamental idea is that frequent direct follows relationships between event labels can represent unusual behavior. These connections have been identified and eliminated from an automaton built from the event log using orientation-based replaying, and the original log is then updated appropriately.

In this work (BPI Challenge 2012), a technique for removing infrequent activities from process running logs was presented. The intention is to substitute follows relationships be- tween event labels for uncommon behavior. In addition to commonly used process identification techniques, our team utilized a variety of simulated and real-world logs to show the effectiveness and performance of the proposed system. These two components are essential to a finding algorithm's accuracy and evaluation. Eventually, as the number of occurrences in the log grows, a discovery algorithm will emerge.

Therefore, finding a model takes less time when there are fewer occurrences in the log. Modeling temporal data is challenging due to its dynamic nature and complex evolutionary patterns. Because it builds an automaton from the event log and updates it by removing particular occurrences using alignment-based replay, our approach is successful

REFERENCES

[1]. B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 1998, pp. 80–86.

[2]. Z. Xing, J. Pei, and E. Keogh, "A brief survey on sequence classification," ACM SIGKDD Explorations Newslett., vol. 12, no. 1,pp. 40–48, 2010.

[3]. C. Zhou, B. Cule, and B. Goethals, "Itemset based sequence classification," in Ma- chine Learning and Knowledge Discovery in Databases.New York, NY, USA: Springer, 2013, pp. 353–368.

[4]. N. Lesh, M. J. Zaki, and M. Ogihara, "Scalable feature mining for sequential data," IEEE Intell. Syst., vol. 15, no. 2, pp. 48–56, Mar./Apr. 2000

DOI: 10.48175/IJARSCT-12057

[5]. B. Cule, B. Goethals, and C. Robardet, “A new constraint for mining sets in se- quences,” in Proc. SIAM Int. Conf. Data Mining, 2009, pp. 317

[6]. J. Wang and G. Karypis, “Harmony: Efficiently mining the best rules for classifica- tion,” in Proc. SIAM Int. Conf. Data Mining, 2005,pp. 205–216.

[7]. V. S. Tseng and C.-H. Lee, “Effective temporal data classification by integrating sequential pattern mining and probabilistic induction,” Expert Syst. Appl., vol. 36, no. 5, pp. 9524–9532, 2009.

[8]. W.M.P. van der Aalst. Process Mining - Data Science in Action, Volume 2nd Edi- tion.Springer, 2016

[9]. J. Wang, S. Song, X. Lin, X. Zhu, and J. Pei. Cleaning structured event logs: A graph repair approach. In Proc. of ICDE, pages 3041, 2015.

[10]. W.M.P. van der Aalst, T. Weijters, and L. Maruster. Workflow mining: Discovering process models from event logs. IEEE TKDE, 16(9):11281142, 2004.

[11]. A. Adriansyah, B.F. van Dongen, and W.M.P. van der Aalst. Conformance checking using cost-based fitness analysis. In Proc. of EDOC, pages 5564, 2011.

[12]. R. Conforti, M. Dumas, L. Garca-Banuelos, and M. La Rosa. Beyond tasks and gate- ways: Discovering BPMN models with subprocesses, boundary events, and activity markers. In Proc. of BPM, pages 101117, 2014.

[13]. S.J.J. Leemans, D. Fahland, and W.M.P. van der Aalst. Discovering block-structured process models from event logs containing infrequent behavior. In Proc. of BPM Work- shops, pages 6678, 2014.

[14]. C.C. Aggarwal. Outlier Analysis. Springer, 2013.

[15]. M. Gupta, J. Gao, C.C. Aggarwal, and J. Han. Outlier detection for temporal data: A survey. IEEE TKDE, 26(9):22502267, 2014.

[16]. A. Adriansyah, J. Munoz-Gama, J. Carmona, B.F. van Dongen, and W.M.P. van der Aalst. Alignment based precision checking. In Proc. of BPM Workshops, pages 137149, 2012.

[17]. R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proc. of IJCAI, pages 11371145, 1995.

[18]. W.M.P. van der Aalst, A. Adriansyah, and B.F. van Dongen. Replaying history on process models for conformance checking and performance analysis. Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery, 2(2):182192, 2012.

[19]. M. Gupta, A. Mallya, S. Roy, J.H.D. Cho, and J. Han. Local Learning for Mining Outlier Subgraphs from Network Datasets, pages 7381. 2014.

[20]. Raaele Conforti, Marcello La Rosa, and Arthur H.M. terHofstede Filtering out Infre- quent Behavior from Business Process Event Logs. Transactions on Knowledge and Data Engineering IEEE Feb 2017.

[21]. C.C. Aggarwal. Outlier Analysis. Springer, 2013.

[22]. M. Gupta, J. Gao, C.C. Aggarwal, and J. Han. Outlier detection for temporal data: A survey. IEEE TKDE, 26(9):22502267, 2014.

[23]. M. Gupta, A. Mallya, S. Roy, J.H.D. Cho, and J. Han. Local Learning for Mining Outlier Subgraphs from Network Datasets pages 7381. 2014.

[24]. W.M.P. van der Aalst. Process Mining - Data Science in Action, volume 2nd Edition. Springer, 2016.

[25]. Tax, Niek, Natalia Sidorova, Wil MP van der Aalst, and ReinderHaakma. "Heuristic ap- proaches for generating local process models through log projections." In Compu- tational Intelligence (SSCI), 2016 IEEE Symposium Series on, pp. 1-8. IEEE, 2016.

[26]. Werner, Michael. "Financial process mining-Accounting data structure dependent control flow inference." International Journal of Accounting Information Systems 25 (2017): 57- 80.

[27]. AlShathry, Omar. "Process Mining as a Business Process Discovery Technique." Com- puter Engineering Information Technology 2016 (2016).

[28]. Xia, Xiaoxu, Wei Song, Fangfei Chen, Xuansong Li, and Pengcheng Zhang. "Ea: a proM plugin for recovering event logs." In Proceedings of the 8th Asia-Pacific Sym- posium on Internetware, pp. 108-111. ACM, 2016

DOI: 10.48175/IJARSCT-12057