Classification of Time-Interval and Hybrid Sequential Temporal Patterns

Mohammed GH. I. AL Zamil
Department of Computer Information Systems
Yarmouk University
Irbed, Jordan
Mohammedz@yu.edu.jo

***Abstract*— Due to the rapid growth of information systems that manage temporal data, efficient and automated classification techniques are of great importance. For instance, timely and accessible temporal data enhances critical financial operations such as predicting future stock prices. Similarly, in medical domain, classifying temporal data, which is relevant to patients or critical operations, leads to efficient control and recovery from severe problems. Therefore, time is an essential dimension to many domain-specific problems. This research introduces Temporal-ROLEX; a framework to categorize temporal data that effectively induces semantic temporal patterns. This paper presents an efficient rule-based classification approach for categorizing temporal data. The contributions of this research are 1) formulating Semantic Temporal patterns as a basic classification features, and 2) introducing an induction technique to discriminate semantic temporal patterns. The proposed framework extends ROLEX-SP approach to handle the classification of temporal data in different domains. To illustrate the design, the article provides a detailed mathematical description that relies on set-theory to model the framework of Temporal-ROLEX. Furthermore, this paper provides a detailed description of proposed algorithms to facilitate implementing and reproducing the results. To evaluate the effectiveness of the Temporal-ROLEX, we performed extensive experiments on a weather temporal dataset. Also, the F-measure and support values on weather dataset are reported as well as a scalability and sensitivity analysis to assess the capability of Temporal-ROLEX to work with temporal datasets. Findings indicate a significant improvement of Temporal-ROLEX over some existing techniques. Specifically, Temporal-ROLEX achieves significant enhancement using sequential temporal pattern over existing state-of-the-art techniques. On the other hand, Temporal-ROLEX achieves average performance using hybrid temporal patterns. Finally, the results have been analyzed and justified the factors that affect the performance in both cases.**

Keywords-Temporal Data Analysis; Classification of Temporal Data; Lexical Patterns.

I. INTRODUCTION

Due to the rapid growth of information systems that manage temporal data, efficient and automated classification techniques of temporal data are of great importance. Time is an essential dimension to many domain-specific problems such as financial and medical domains. This paper presents an efficient rule-based classification approach for categorizing temporal data. The contributions of this research are 1) formulating Semantic Temporal patterns as a basic classification features, and 2) introducing an induction technique to discriminate semantic temporal patterns.

ROLEX-SP has been introduced by M. ALZamil and A. Can [1] to categorize domain specific knowledge using specialized rule-based induction and learning methods to produce efficient classification of domain specific knowledge. ROLEX-SP automates the induction and the learning processes by extracting lexical patterns and constructing specialized form of association rules. Such technique handles the problems of multiclass classification and feature imbalance problems. This research introduces Temporal-ROLEX; a framework to categorize temporal data that effectively induces semantic temporal patterns

Temporal-ROLEX is intended to find temporal relationships such as: during, after, overlap, start, finish and equal. However, it defines a form of association rules that generate not only efficient patterns to classify events, but also minimize the margin error to enhance the overall performance of the classification task. Figure 1 shows during temporal event, in which event e1 starts and finishes during the execution of event e2. The work in this paper is restricted to during relationship, since it is able to represent hybrid relations among temporal events. In other words, a hybrid relationship is able to describe before and after relationships.

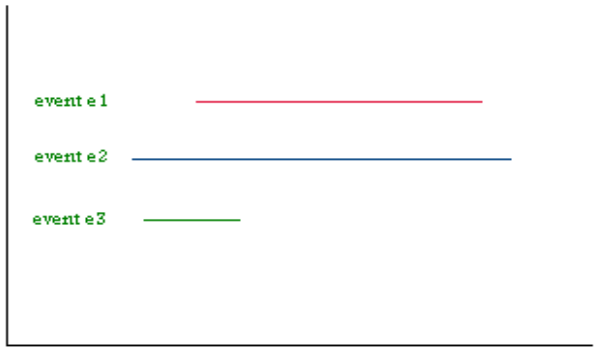

Figure 1. During sequential pattern (i.e., events e1 starts and finishes during the execution of e2)

 ISBN: 978-1-61208-227-1

This paper is structured as follows: Section I introduces the problem understudy and proposes the direction of the research solution. Section II provides a literature review of strictly related work of other researchers and compares them with the proposed method. Section III defines the formal model of the proposed methodology as a background to theoretical and empirical work. Section IV details the framework that is followed to classify temporal data and the algorithm that has been applied to produce the empirical results. Section V provides a description of the experimental setup. Finally, Section VI discusses the conclusion of this work and justifies the results.

II. RELATED WORK

Attempts have been made to construct temporal features in order to construct association rules such as those discussed in Bruno and Garza [2], Miao et al. [3], and Chiang et al. [4]. In Bruno and Garza [2], association rules have been developed to cope with outlier detection using functional quasi dependency. The technique does not model time-delay as a part of association rules. The delta function that associate two temporal attributes X and Y assumes no delay. The technique in Bruno and Garza [2] handled time-delay explicitly, which affect the overall performance as well as efficiency of the classification process; which is not crucial in outlier detection task.

Chiang et al. [4] have proposed a mathematical model to extract temporal patterns to track customer buying habits. The model is developed to capture temporal characteristics of business data in single-point-of-time events. Our proposed methodology focuses on time intervals as well as single point of time events. Similarly, our proposed technique benefits from the formal definition in [4], in that we formulate the temporal patterns using similar mathematical aspects.

Zhang et al. [5] have proposed a method to extract *during* temporal patterns. A during temporal pattern (DTP) is a special case of interval temporal patterns. These patterns provide valuable information in broadcasting future information such as weather and stock broadcasting. Kong et al. [6] have presented the notion of multi temporal patterns using predicates: before, during, equal and overlap.

Winarko and Roddick [7] have Introduced ARMADA, an algorithm to discover interval time temporal rules. ARMADA research asserts that time-stamps relationships such as *during* could be more useful than solid time interval. Classifying time-interval events into temporal clusters provide meaningful information in different application areas such as financial analysis and weather broadcasting. Unlike ARMADA association rules, our work relies on discovering hybrid temporal rules that could be represented using *during* relation.

Although performance plays a significant role in assessing classification techniques, pre-processing tasks might be crucial in many applications in terms of scalability and efficiency. However, temporal datasets dimensions are characterized as huge ones. Techniques to reduce such dimensionality are important to produce scalable temporal mining systems. We applied methods in Stacey and McGregor [8] and Wang and Megalooikonomou [9] to reduce the dimensionality of time series.

In the literature, there are many data mining and knowledge discovery techniques on medical domain and biomedical data [10, 11, 12, 13]. The contribution of this article over existing ones is the ability of Temporal ROLEX to handle timely information regardless of its domain. Further, the method proposed in this article deals with time as a classification feature. The later might negatively affect the classification performance, but it adds the advantage of enhancing timely information classification

III. BACKGROUND

For the purpose of defining the formal model of the temporal classification problem, Inductive-Logic-Programming (ILP) [14] is used as follows: given

1. A finite set TC of unrelated temporal classes of the form $\{Tc_1, Tc_2, ..., Tc_k\}$ where $k > 1$, meaning that there are many temporal classes and the assigned label of a class do not affect the labeling of other classes. For instance, an event might be classified under *during* class and *overlap* class at the same time if this event belongs to different set of events.
2. A set $E = \{e_1, e_2, \ldots, e_n\}$ of events such that $\forall(j)\exists(\ Q \subseteq TC \wedge |\ Q\ | = v) : e_j \in Q$ where $1 \le v \le k$ and $1 \le j \le N$, meaning that an event might belong to more than one temporal class; Q is a subset of the set of temporal classes.
3. A set of states $S = \{s_1, s_2, \ldots, s_m\}$ each of which represents a state of the current environment such as: *raining and shining* in the weather dataset.
4. A set of time-intervals $T = \{t_1, t_2, \ldots, t_n\}$, where $t_i = \{st_i, et_i\}$ represents the start and end time of a given event e_i
5. A set P_{ci}^{+} of positive patterns consisting of atomic facts of the form $p_{ci}^{+} \in E_{Tci}$ such that $(p_{ci}^{+} \in e \wedge e \in E_{Tci}) \Rightarrow e \in Tci$; a positive pattern under class Tci that occurs in the subset E_{Tci}, which represent a set of events that belong to the class Tci.
6. A set P_i^{-} of negative data patterns; patterns that represent an event but does not refer to

class Tci. In other words, they represent outliers or rare cases.

7. The function $[g(a_\alpha) = \{ e_1(a_\alpha, t_1), e_2(a_\alpha, t_2), \ldots, e_k(a_\alpha, t_k)\}]$ includes all the interval times in which the state a_α occurs.

construct a classifier H_{ci} that consists of all positive and negative facts. In other words, the classifier represents a set of association rules to forecasting a temporal class or a set of temporal classes of a given set of events based on the presence or absence of some facts in that set.

The learning task of Temporal-ROLEX generates association rules such that: given a category $Tc_i \in TC$, a positive pattern $p^+_{Tci} \in P_{Tci}{}^+$ associates with class Tci, and a set of negative patterns $P_i^- (P^- \cap P^+ = \phi)$, where P^- is the set of all negative patterns and P^+ is the set of positive patterns, the classifier H_{Tci} of class Tci is defined as a set of rules. We used the rule's representation in [15] as follows:

$$\left[Tc_i \leftarrow p^+_{Tci} \in g(a_\alpha),\ \neg(p^-_{i1} \in g(a_\alpha)) \wedge \neg(p^-_{i2} \in g(a_\alpha)) \wedge \ldots \wedge \neg(p^-_{im} \in g(a_\alpha))\right]$$

If a positive example p^+_{ci} occurs in document $g(a_\alpha)$ and none of the negative patterns occur in $g(a_\alpha)$, the classifier will assign event e under class Tci. Notice that, negative patterns are prevented from undoing the effect of other categories' positive ones.

IV. FRAMEWORK

Let $e_j = \{s_i, t_j\}$ and $e_k = \{a_l, t_k\}$ be two events in the temporal dataset. Both e_j and e_k are called during events if e_j has executed during the execution of e_k. For any two given states a_i and a_k, a_i is called to be during a_k denoted as $a_i \Rightarrow^d a_k$. Our goal is to define a set of positive and negative predicates to predict during temporal patterns.

Instead of the accuracy formula that has been applied in the previous version of ROLEX-SP, the function support that has been defined in [16] has been used to induce positive and negative patterns as well. Given $|g(a_\alpha)|$, the number of the time intervals included in all instances (records in the dataset) of a_α, the maximum number of time intervals among all states $|g_0|$:

$$Support(a_\alpha) = \frac{|g(a_\alpha)|}{|g_0|} \quad (1)$$

It represents the relative frequency of time intervals for a given state with respect to the number of time intervals for a most frequent state.

The proposed induction algorithm is shown in Figure 2.

LSP_Generator

- **Goal**: to extract positive and negative syntactic patterns from the set TS
- **Input**: TS, C.
  - **TS**: the set of labeled training set of data records
  - **C**: the set $C = \{c_1, c_2, \ldots c_k\}$ of categories where $k > 1$.
- **Output**: P^+, P^-
  - $P^+ = (P^+_{c1}, P^+_{c2}, \cdots, P^+_{ck})$ A set of positive patterns relevant to all categories.
  - $P^- = (P^-_1, P^-_2, \cdots, P^-_k)$ A set of negative patterns relevant to all categories.
- **Method**: Apply the following instructions

```
Begin
1     P+ = (), P- = ()
2     For each ci ∈ C
3       For each d ∈ TSc
4         P = parse(d, Rec(ci))
5         For each p ∈ P
6               Support(a_α) = |g(a_α)| / |g_0|
                if Support(a_α) ≥ threshold then
7                 P+_ci = P+_ci ∧ p
                Else
                  P-_i = P-_i ∧ p
8         Next p
9       Next d
10    Next ci
11    return(P+, P-)
End
```

Figure 2. Induction Algorithm

V. EXPERIMENT AND ANALYSIS

In this section, the results have been collected from applying Temporal-ROLEX on a weather dataset. The data set has been obtained from a weather station in Jordan in 2009. The dataset consists of 14 attributes: wind direction, average wind speed, maximum wind gust, average hourly temperature, percentage relative humidity, global hourly radiation, hourly sunshine duration, hourly precipitation duration, hourly precipitation amount, horizontal visibility, fog, snow, etc. Most of the collected values were continuous. To proceed, the pre-processing techniques to discriminate and convert the records into temporal ones have been applied, which are consisting of event name, start time, end time, and state.

A. F-Measure

First, we compute the recall and precision values relevant to every category according to the following formulas:

 ISBN: 978-1-61208-227-1

$$\mathrm{Pr} = \sum_{c \in C} |TP_c| / \sum_{c \in C} |TP_c| + |FP_c| \qquad (2)$$

$$\mathrm{Re} = \sum_{c \in C} |TP_c| / \sum_{c \in C} |TP_c| + |FN_c| \qquad (3)$$

where $|TP_c|$ is the number of correctly classified records in the testing set under category c, $|FP_c|$ is the number of incorrectly classified records in the testing set under category c, and $|FN_c|$ is the number of records in the testing set, which were not classified under category c but should have been. The F-measure is defined as follows:

$$F = \mathrm{Pr} \times \mathrm{Re} / (1-\alpha)\,\mathrm{Pr} + \alpha\,\mathrm{Re} \qquad (4)$$

where $\alpha \in [0,1]$

$$Average\ F_{macro} = \sum_{i=1}^{|C|} \frac{F_i}{|C|} \qquad (5)$$

where |C| is the number of categories in the dataset.

The results indicate that Temporal-ROLEX achieves 67.8% average F-Measure. The experiments show that Temporal-ROLEX achieves significant enhancement using sequential temporal pattern over existing state-of-the-art techniques on the same dataset such as DTP [16] that achieve 66.2%. On the other hand, Temporal-ROLEX achieves average performance using hybrid temporal patterns; i.e., 63.7 while DTP achieve 65.1%. The results have been analyzed and justified the factors that affect the performance in both cases

B. *Sensitivity Analysis*

In order to evaluate the results, the analysis task considers four sensitivity attributes, which measure the quality of empirical results. These attributes include the number of valid patterns, the effect of rules on average F-measure, the execution time versus number of rules, and the execution time versus the number of events.

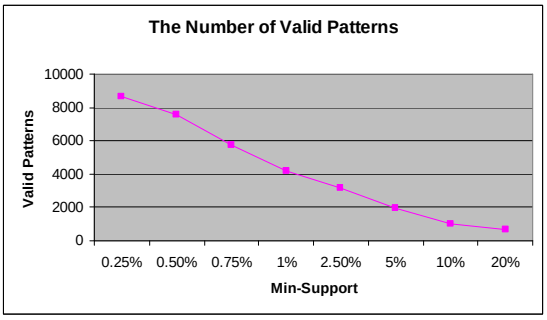

Figure 3. The Number of valid Patterns and Minimum Support

Figure 3 shows that Temporal-ROLEX performs well at low percentage of support measure. In other words, the rules induced using the proposed induction algorithm are able to classify valid patterns correctly among low number of time intervals in the training set.

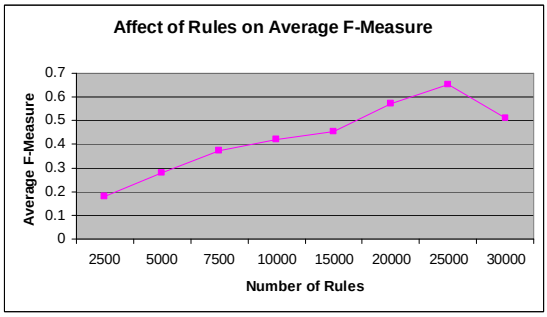

Figure 4. The Number of Rules and The average F-Measure

Figure 4 concludes a positive relationship between f-measure and the number of rules; the higher the number of rules, the higher the f-measure. This property demonstrates that in order to achieve high performance, the induction algorithm has to be fed with large training set to produce rules that cover all, or at least, most patterns.

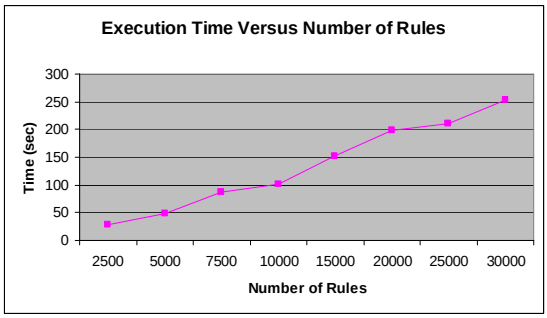

Figure 5. The Execution Time and The Number of Rules

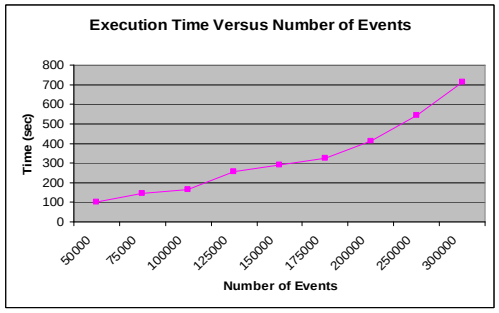

Figure 6. The Execution Time and Number of Events

Finally, Figures 5 and 6 show that the execution time increased as the number of rules and/or events increased.

 ISBN: 978-1-61208-227-1

VI. CONCLUSION

This paper presented a rule-based method for categorizing temporal records. The contributions of this research are 1) formulating Semantic Temporal patterns as a basic classification features, and 2) introducing an induction technique to discriminate semantic temporal patterns. Experiments have been performed on a weather dataset in order to evaluate the proposed method and compare our work with well known algorithms in the literature. Findings indicate a significant improvement of Temporal-ROLEX over a well known technique; DTP. Specifically, Temporal-ROLEX achieve significant enhancement using sequential temporal pattern. On the other hand, Temporal-ROLEX achieves average performance using hybrid temporal patterns.

Furthermore, Temporal-ROLEX achieved statistically significant improvement. Applying syntactic patterns, both positive and negative, enhances the accuracy of Temporal-ROLEX over the other method.

The article also provided a sensitivity analysis to the performance of Temporal-ROLEX as a function to the number of association rules and the number of data elements in the training set. The results indicated that Temporal-ROLEX was affected by the number of rules positively. On the other hand, the observations during experiments indicated that the number of records in the training set does not affect the overall performance of the learning process.

REFERENCES

[1] Al Zamil, M. and Betin-Can, A. ROLEX-SP: Rules of lexical syntactic patterns for free text categorization. Knowledge-Based Systems 24 (2011) 58–65.

[2] Bruno, G. and Garza, P. TOD: Temporal outlier detection by using quasi-functional temporal dependencies. Data & Knowledge Engineering 69 (2010) 619–639

[3] Miao, O., Li, O., and Dai, R. AMAZING: A sentiment mining and retrieval system. Expert Systems with Applications 36 (2009) 7192–7198.

[4] Chiang, D., Wang, Y., and Chen, S. Analysis on repeat-buying patterns. Knowledge-Based Systems 23 (2010) 757–768.

[5] Zhang, L., Chen, G., Brijs, T. and Zhang, X. Discovering during-temporal patterns (DTPs) in large temporal databases. Expert Systems with Applications 34 (2008) 1178–1189.

[6] Kong, X., Wei, O. and Chen, G. An approach to discovering multi-temporal patterns and its application to financial databases. Information Sciences 180 (2010) 873–885.

[7] Winarko, E. and Roddick, J. ARMADA – An algorithm for discovering richer relative temporal association rules from interval-based data. Data & Knowledge Engineering 63 (2007) 76–90.

[8] Stacey, M. and McGregor, C. Temporal abstraction in intelligent clinical data analysis: A survey. Artificial Intelligence in Medicine (2007) 39, 1—24.

[9] Wang O. and Megalooikonomou, V. A dimensionality reduction technique for efficient time series similarity analysis. Information Systems 33 (2008) 115–132

[10] Simonic, K. M., Holzinger, A., Bloice, M. & Hermann, J. (2011) Optimizing Long-Term Treatment of Rheumatoid Arthritis with Systematic Documentation. Proceedings of Pervasive Health - 5th International Conference on Pervasive Computing Technologies for Healthcare. Dublin, IEEE, 550-554

[11] Holzinger, A., Simonic, K. M., andYildirim, P. (2012) Disease-disease relationships for rheumatic diseases. COMPSAC 2012. Izmir,Turkey

[12] Holzinger, A., Scherer, R., Seeber, M., Wagner, J., and Müller-Putz, G. (2012) Computational Sensemaking on Examples of Knowledge Discovery from Neuroscience Data: Towards Enhancing Stroke Rehabilitation. In: Böhm, C. (Ed.) International Conference on Information Technology in Bio- and Medical Informatics - ITBAM 2012 Heidelberg, Berlin, New York, Springer, 166-168

[13] Holzinger, A. (2012). On Knowledge Discovery and interactive intelligent visualization of biomedical data: Challenges in Human–Computer Interaction & Biomedical Informatics. DATA - International Conference on Data Technologies and Applications, Rome-Italy.

[14] Lavrac, N. and Dzeroski, S. Inductive Logic Programming: Techniques and Applications. Ellis Horwood, New York (1994).

[15] P. Rullo, V. Policicchio, C. Cumbo, S. Iiritano, Olex: Effective rule learning for text categorization, IEEE Transactions on Knowledge and Data Engineering 21 (8) (2009) 1118‑1132.

[16] Wu, S. and Chen, Y. Discovering hybrid temporal patterns for interval-based events, IEEE Transactions on Knowledge and Data Engineering 19 (6) (2007) 742-758.

 ISBN: 978-1-61208-227-1