TLEX: A Temporal Analysis Tool for Time Series Data

Mohammed AL Zamil
Department of Computer Information Systems
Yarmouk University
Irbed, Jordan, 21163
Email: Mohammedz@yu.edu.jo

Bilal Abu AL Huda
Department of Management Information Systems
Yarmouk University
Irbed, Jordan, 21163
Email: abul-huda@yu.edu.jo

***Abstract*—Time is an essential dimension to many domain-specific problems, such as the medical and financial domains. This research introduces TLEX (Temporal Lexical Patterns), a framework to categorize temporal data that effectively induces semantic temporal patterns. TLEX is a rule-based classification framework dedicated to enhance the classification accuracy by focusing on eliminating outliers and minimizing classification errors. The contributions of this research are 1) formulating semantic temporal patterns as basic classification features, and 2) introducing an induction technique to discriminate semantic temporal patterns. To illustrate the design, the paper provides a detailed mathematical description that relies on set-theory to model the framework of TLEX. Furthermore, a detailed description of the proposed algorithms to facilitate implementing and reproducing the results has been described. Further, to evaluate the effectiveness of TLEX, extensive experiments have been performed on a weather temporal dataset. Accordingly, the F-measure and support values on weather dataset have been reported. Further, a sensitivity analysis to assess the capability of TLEX to work with temporal datasets has been provided. The findings indicate a significant improvement of Temporal-ROLEX over some existing techniques.**

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

I. INTRODUCTION

Time is an essential dimension to many domain-specific systems such as financial and medical data analysis. However, temporal data mining is concerned with such analysis in the case of ordered data records with temporal interdependencies. During the last decade, many interesting techniques of temporal data mining were proposed and shown to be useful in many applications areas. Since temporal data mining brings together techniques from different fields, such as statistics, machine learning and databases, the literature is scattered among many different sources.

Temporal data mining is commonly concerned with data mining of large sequential datasets that have been ordered chronologically with respect to some index. For example, time series constitute a popular class of sequential data, where records are indexed by time. Other examples of sequential data could be text, gene sequences, protein sequences, lists of moves in a chess game, etc. Here, although there is no notion of time as such, the ordering among the records is very important and is central to the data description and modeling.

Consider the temporal relation among three customers whose corresponding transaction sequences are as follows:

Cust. 1. $[\{X_1\ X_2\ \},\{X_3\ X_1\ X_4\ \},\{X_2\ X_5\}]$
Cust. 2. $[\{X_5\ \},\ \{X_1\ X_2\}]$
Cust. 3. $[\{X_1\ \},\{X_1\ X_2\ X_5\ X_6\ \}]$

where $\{X_i\}$ represents the items bought in a single transaction. For instance, customer 1 made 3 visits to the market. In her first visit, she bought 2 items $\{X_1\ X_2\}$. In her second and third visits, she bought 3 items $\{X_3\ X_1\ X_4\}$ and 2 items, $\{X_2\ X_5\}$ respectively. Temporal patterns are frequent sequences of actions that could be useful for analyzing data and predicting futures. In the above example, we can extract relations such as the sequence $[\{X_5\ \},\ \{X_1\ X_2\}]$ contained in $[\{X_1\ X_2\ \},\{X_3\ X_1\ X_4\ \},\{X_2\ X_5\ \}\ ]$ but not in $[\{X_1\ \},\{X_1\ X_2\ X_5\ X_6\ \}]$.

This paper presents an efficient rule-based classification approach, called TLEX, 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.

TLEX framework relies on the formal definition of ROLEX-SP that has been introduced by M. AL Zamil and A. Can [1] to classify medical knowledge using a dedicated rule-based induction and learning techniques to come with efficient classification of domain knowledge. ROLEX-SP automates the induction and the learning processes by extracting textual patterns and building a specialized form of association rules. Such technique handles the problems of multiclass classification and feature imbalance problems.

To illustrate, consider the example of raining phenomena. A person might hear the thunder during and after its occurrence. Duration represents the persistence of an event over many time points. Also, rain and thunder might occur

 ISBN: 978-1-61208-552-4

concurrently without a particular order. Finally, some events might intersect each other such as sun shining, raining and rising of the rainbow.

Section II discusses the related work. Section III provides background definitions of formalities that have been applied in this work. Section IV defines the temporal model as well as the form of the temporal pattern. Section V discusses the methodology for creating our proposed classifier. Section VI discusses the experiments and findings. Finally, Section VII concludes with a brief discussion of findings.

II. RELATED WORK

Early works on mining temporal patterns rely on Apriori-style approaches such as the algorithm in [2], in which a breadth-first search strategy is applied to compute the support of item sets. As this strategy is not efficient on large datasets in terms of accuracy [3], recently efficient algorithms based on unsupervised elicitations of temporal relations have been proposed for the purpose of enhancing the performance of temporal classifiers. H-DFS [4] and KarmaLego [5] are based on an enumeration tree structure to classify temporal information. The implementation of enumerated decision trees supports the accuracy of classification results in an unsupervised manner.

Winarko and Roddick [6] have introduced ARMADA, which is an algorithm to discover interval time temporal rules. Recent work in this field asserts that time-stamps relationships such as the "during" relation could be more useful than solid time interval. Unlike ARMADA association rules, our work relies on discovering hybrid temporal rules that could be represented using during relation. In [7], TPrefixSpan has been introduced to apply prefix span technique using interval boundaries pruning patterns. Also, IEMiner [8] has implemented a prior based strategy with counters and pruning to improve the accuracy of IEClassifier; that is used to classify temporal sequential records.

Attempts have been made to construct temporal features in order to construct association rules such as those discussed in Bruno and Garza [9], Miao et al. [10], and Chiang et al. [11]. In Bruno and Garza [9], 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 technique in Bruno and Garza [9] handled time-delay explicitly which affects the overall performance as well as efficiency of the classification process, which is not crucial in outlier detection task.

Chiang et al. [11] have proposed a mathematical model to extract temporal patterns to track customer buying habits. 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 Chiang et al. [11] in that we formulate the temporal patterns using similar mathematical aspects. Zhang et al. [12] have proposed a method to extract during temporal patterns DTP. DTP is a special case of interval temporal patterns. Kong et al. [13] have presented the notion of multi temporal patterns using predicates: before, during, equal and overlap.

Temporal datasets dimensions are characterized as huge ones. Techniques to reduce such dimensionality are important to produce scalable temporal mining systems. The proposed technique applied methods in Stacey and McGregor [14] and Wang and Megalooikonomou [15] to reduce the dimensionality of time series.

III. BACKGROUND

Definition 1 (Temporal Sequence):

A temporal sequence is a chronologically ordered set of events of the form:

$$TSq = \{e_1(st, et), e_2(st, et), ..., e_n(st, et)\}$$

where (st, et) is a nonempty set in which sd is the start-time of an event and ed is the end-time of the same event, i.e., $st, et \in TIME$.

Definition 2 (Temporal Sequence Similarity):

Two temporal sequences are said to be similar, i.e. $Sim(TSq_1, TSq_2)$, if and only if all the following conditions hold:

1. $\forall(i)\{e_i(TSq_1) = e_i(TSq_2)\}$
2. $\forall(i)\{Len[e_i(TSq_1)] = Len[e_i(TSq_2)]$
3. $TSq1 \wedge TSq2$ are not empty

where $Len[(e_i(Sq)]$ refers to the event duration (i.e. $et - st$). This relation is useful to eliminate redundancy resulted from later categorization process.

Definition 3 (Temporal Sequence Dissimilarity):

Two temporal sequences are said to be dissimilar, i.e. $Dis(TSq_1, TSq_2)$, if $\neg Sim(TSq_1, TSq_2)$ holds.

Definition 5 (Support of Temporal Sequence):

The function $Supp(Tsq, D) \leftarrow \{Si \mid Si \in D \wedge Tsq \in Si\}$ is used to determine the support of a frequent pattern Tsq in a given dataset D , in which Si is a sequence fragment. Xingzhi et al. [16] illustrate the application of support model in data mining by theorem proving and case studies.

IV. TEMPORAL CLASSIFICATION

The temporal classification problem is defined according to the learning description of ILP (Inductive Logic Programming) Lavrac and Dzeroski [17], as follows: given

1. A finite set TC of independent temporal classes of the form $\{Tc_1, Tc_2, ..., Tc_k\}$ where $k > 1$, meaning that there are many temporal classes and the classification results of a class do not affect the classification results of other classes.

2. A set $E = \{e_1, e_2, \ldots, e_n\}$ of events such that $\forall(j)\exists(\ Tsqi \subseteq TC \wedge |\ Tsqi\ | = v) : e_j \in Tsqi$ where $1 \le v \le k$ and $1 \le j \le N$, meaning that an event might belong to more than one temporal class; Si 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 ground logical 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 class Tci.

6. A set P_i^{-} of negative facts; 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.

A classifier H_{ci} should be consistent with all positive and negative patterns. In other words, the classifier is a set of association rules to predict a temporal class or a set of temporal classes of a given set of events based on the presence or absence of some patterns in that set.

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.

V. CLASSIFICATION METHODOLOGY

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 accuracy formula that has been applied in the previous version of ROLEX-SP, the function *support* that has been defined in [19, 20, 21] 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 sates $|g_0|$:

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

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. LSP Generator, Figure 1, implements our proposed methodology to induce classification rules.

LSP Generator

Goal: to extract positive and negative syntactic patterns from the set TS
Input: TS, C.

Output: P^{+}, P^{-}
Method: Apply the following instructions

Begin
1 $P^{+} = \{\}, P^{-} = \{\}$
2 For each $ci \in C$
3 For each $d \in TSc$
4 $P = parse(d, \mathrm{Re}\, c(ci))$
5 For each $p \in P$
6 $Support(a_\alpha) = \frac{|g(a_\alpha)|}{|g_0|}$
if $Support(a_\alpha) \ge threshold$ *then*
7 $P_{ci}^{+} = P_{ci}^{+} \wedge p$
Elseif
$P_i^{-} = P_i^{-} \wedge p$
8 Next p
9 Next d
10 Next ci
11 return(P^{+}, P^{-})
End

Figure 1 The Relationship between Execution Time and The Number of Rules

The validation is responsible for evaluating the extracted rules and generating a classifier H_{ci}. Therefore, the classifier should contain the best rules to represent an event.

Definition 6 (Representative Set RS): given a set of rules sorted according to its support, RS is the set of rules of the form

$$c \leftarrow p_{ci}^{+} \in d, \quad \neg(p_{i1}^{-} \in d) \wedge \neg(p_{i2}^{-} \in d) \wedge \cdots \wedge \neg(p_{iM}^{-} \in d)$$

that have the highest support. Given a rule R and a set of events $E_{ci} \in C \times E$; a set of events that belong to a specific category, let $n_{covers}(R, ci)$ be the number of events covered by R under category ci , and $| E_{ci} |$ be the number of events in E_{ci} :

$$coverage(R, ci) = n_{covers}(R, ci) / | E_{ci} |$$

Accordingly, the validation phase, then, tries to optimize the problem such as: given $R = \{R_{c1}, R_{c2}, \cdots R_{ck}\}$ where $R_{ci} = \{R_1, R_2, \cdots, R_W\}$ and $w = | P_{ci}^{+} |$, the algorithm is responsible to produce the set $RS_{ci} = \{R_1, R_2, \cdots R_x\}, where\ x \le w\ and\ RS_{ci} \subseteq R_{ci}$, of rules such that: $Coverage(RS_{ci})$ is the maximum.

Definition 7 (Redundant Rule): a rule Rj is a redundant rule if one of the following conditions holds:

1. $(\forall i)(\exists j) : R_i = R_j \wedge i \neq j$

2. $(\forall i)(\exists j) : Coverage(R_j) \subseteq Coverage(R_i) \wedge i \neq j$

Thus, getting rid of redundant rules, which are equal rules or rules that cover the same set of another rules, will enhance the overall performance of the classification task.

VI. EXPERIMENT AND RESULTS

During our experiment, the proposed technique has been applied on a weather dataset. The dataset has been collected from a weather station in Jordan in 2009. The empirical dataset holds 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. The dataset has been processed to discriminate and convert the records into temporal ones consisting of event name, start time, end time, and state. The F-measure achieved during experiments is 81% at minimum support ranges up to 20% as shown in Figure 2

Figure 2 shows that the overall performance of our proposed technique is directly affected by the number of generated rules. Therefore, the higher the number of rules, the better the performance achieved by TLEX.

Figure 3 shows that additional running time is required while increasing the number of rules in our classifier. In fact, the results showed that the required time increased linearly as the number of rules in-creased.

Figure 4, on the other hand, shows that the time required by processing events is much greater than the one for events. However, a linear relation is clear between the running time and the number of events.

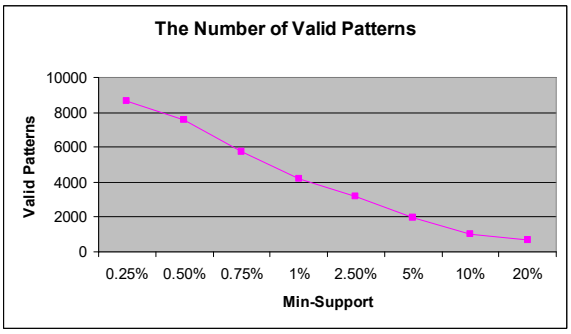

Figure 2 The Relationship between Number of valid Patterns and Minimum Support

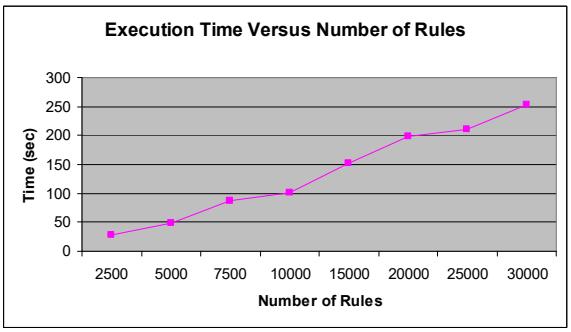

Figure 3 The Relationship between Execution Time and The Number of Rules

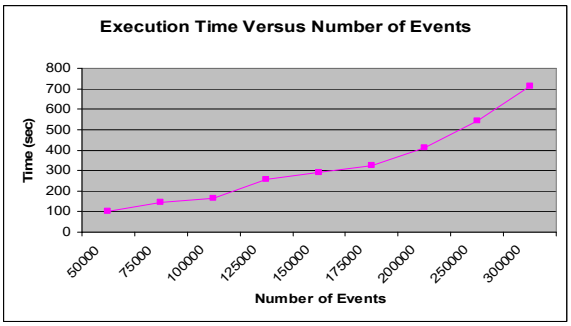

Figure 4 The Relationship between Execution Time and Number of Events

VII. CONCLUSION

In this paper, we 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. We performed experiment on a weather dataset in order to evaluate the proposed method and compare our work with well known algorithms in the literature. Specifically, Temporal-ROLEX achieve significant enhancement using sequential temporal pattern. On the other hand, Temporal-ROLEX achieves average performance using hybrid temporal patterns.

 ISBN: 978-1-61208-552-4

Also, the improvement achieved by Temporal-ROLEX is statistically significant. The use of syntactic patterns, both positive and negative, contributes on increasing the accuracy of Temporal-ROLEX over the other method.

In addition, we also provided a sensitivity analysis to the performance of Temporal-ROLEX as a function to the number of rules and the number of records in the training set. The results indicated that Temporal-ROLEX was positively affected by the number of rules. On the other hand, our 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] M. G. Al Zamil and A. B. Can. ROLEX-SP: Rules of lexical syntactic patterns for free text categorization. Knowledge-Based Systems, 24(1), 2011, pp. 58-65.

[2] R. Agrawal, T. Imieliński, and A. Swami. (1993, June). Mining association rules between sets of items in large databases. In Acm sigmod record (Vol. 22, No. 2, pp. 207-216). ACM.

[3] T. Page, A. L. Heathwaite, L. J. Thompson, L. Pope, and R. Willows. Eliciting fuzzy distributions from experts for ranking conceptual risk model components. Environmental Modelling & Software, 36, 2012, pp. 19-34.

[4] P. Papapetrou, G. Kollios, S. Sclaroff, and D. Gunopulos. Discovering frequent arrangements of temporal intervals. In Data Mining, Fifth IEEE International Conference on, November 2005, pp. 1-8. IEEE.

[5] R. Moskovitch and Y. Shahar. Medical temporal-knowledge discovery via temporal abstraction. In AMIA, 2009, pp. 452–456.

[6] E. Winarko and J. F. Roddick. ARMADA–An algorithm for discovering richer relative temporal association rules from interval-based data. Data & Knowledge Engineering, 63(1), 2007, pp. 76-90.

[7] S. Y. Wu and Y. L. Chen. Mining nonambiguous temporal patterns for interval-based events. IEEE transactions on knowledge and data engineering, 2007, 19(6).

[8] D. Patel, W. Hsu, and M. L. Lee. Mining relationships among interval-based events for classification. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, June 2008, pp. 393-404. ACM.

[9] G. Bruno and P. Garza. TOD: Temporal outlier detection by using quasi-functional temporal dependencies. Data & Knowledge Engineering, 69(6), 2010, pp. 619-639.

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

[11] D. A. Chiang, Y. H. Wang, and S. P. Chen. Analysis on repeat-buying patterns. Knowledge-Based Systems, 23(8), 2010, pp. 757-768.

[12] L. Zhang, G. Chen, T. Brijs, and X. Zhang. Discovering during-temporal patterns (DTPs) in large temporal databases. Expert Systems with Applications, 34(2), 2008, pp. 1178-1189.

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

[14] M. Stacey and C. McGregor. Temporal abstraction in intelligent clinical data analysis: A survey. Artificial intelligence in medicine, 39(1), 2007, pp. 1-24.

[15] Q. Wang and V. Megalooikonomou. A dimensionality reduction technique for efficient time series similarity analysis. Information systems, 33(1), 2008, pp. 115-132.

[16] M. G. Zamil and S. Samarah. Dynamic event classification for intrusion and false alarm detection in vehicular ad hoc networks. International Journal of Information and Communication Technology, 8(2-3), 2016, pp. 140-164.

[17] S. Dzeroski and N. Lavrac. Inductive logic programming: Techniques and applications. 1994.

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

[19] M. G. Zamil and S. Samarah. Dynamic rough-based clustering for vehicular ad-hoc networks. International Journal of Information and Decision Sciences, 7(3), 2015, pp. 265-285.

[20] M. A. Zamil, S. Samarah, A. Saifan,, and I. A. Smadi. Dispersion–based prediction framework for estimating missing values in wireless sensor networks. International Journal of Sensor Networks, 12(3), 2012, pp. 149-159.

[21] S. Samarah, M. A. Zamil, A. Aleroud, M. Rawashdeh, M. Alhamid, and A. Alamri. An Efficient Activity Recognition Framework: Toward Privacy-Sensitive Health Data Sensing. IEEE Access. DOI: 10.1109/ACCESS.2017.2685531. 2017

 ISBN: 978-1-61208-552-4