Multi-Domain Anomalous Temporal Association (Multi-DATA)

Suraksha Shukla, Vandana P. Janeja*
University of Maryland Baltimore County, Baltimore, USA
{shuklas1,vjaneja}@umbc.edu

ABSTRACT

Temporal data from a sensor in a sensor network can capture knowledge for example, weather trends, and precipitation levels in a region over time. Traditional temporal data mining has looked at patterns, such as anomalies, in each temporal data stream. However, in many cases one temporal stream may not provide clear understanding of the phenomena at work. For example, measurement of solar radiation may have relationships to temperature, humidity levels or populations. To study real world phenomena and inter relationships between different temporal streams, in this paper, we propose a novel approach to discover the temporal relations between multiple distinct domains represented by multiple distinct temporal data collected at a location. Different types of sensors or sensors monitoring different types of measures can be considered as distinct domains. In some cases, even the same data may be measuring different types of behaviors. Our goal is to discover the relationship between distinct domains using interesting temporal events in them. These interesting temporal events are mined using traditional temporal anomaly detection methods. In addition, relations between two application domains are not always simple since there can be some time-delay in these relationships. Thus, focusing on relations found using intersecting time events alone is not sufficient. To address this we employ the concept of not only direct overlap but also proximity between temporal events across domains to find the direct and time-delayed relationships. Performing a multi domain analysis can help analysts move towards notions of explainability in a complex phenomena environment, which essentially mimics the real world. We have achieved optimistic results in our experiment on multiple datasets with verified ground truth.

KEYWORDS

Multi-domain, anomaly detection, data heterogeneity, temporal overlaps in anomalies, delayed correlation

1 Introduction

Traditional temporal anomaly detection techniques identify the anomalous patterns in a single time series data. However, detected unusual behavior in one time series can have impacts on other variables as well [1] and analyzing time series data as an independent feature cannot identify the complex nature of real world problems. In addition, anomalies in one domain are generally impacted by other application domains. In majority of the cases, an observed phenomenon in one domain can very well be explained by linking other domain data. Considering the impact and quantifying the level of impact of other domains leads to a more accurate analysis and result while also revealing possible explanations of the observed behavior. Such multi-domain associations generally identify a potential hypothesis, which needs to be further investigated for ground truth and validation.

In this paper, our goal is to discover Multi-Domain Anomalous Temporal Associations (Multi-DATA). In our approach, we analyze the discovered anomalies in individual domains for two conditions, first we identify if there is an overlap between anomalous time sequences across domains, and second, if there is no direct overlap, we identify if the anomalies from different domains are within the specified proximity. In this latter case, we measure the time-delayed correlation. For both cases, we use association rule mining to discover the relation between those domains.

Multi-DATA analyzes complex connections across disparate domains. Finding data, especially with ground truth is a challenge in itself. Once found, it requires rigorous data cleaning and transformations. As we are using multiple domains, we also have to deal with data heterogeneity across domains. Thus, discovery of such associations across multiple domains needs a framework that does a comprehensive analysis to capture all possible cases of temporal relations as simultaneous impacts and delayed impacts.

2 Related Work

Discovering hidden relations between sequences and subsequences of events is the goal of temporal data mining [2]. Roddick et al. (2001) [3] used the Apriori-like method, causal rule on temporal data to discover rules comprising time information. As compared to conventional association rule mining, temporal association rule adds time information which might be a time point or time range [4]. Episodal association [5] discovers periodic occurrence of interesting events. Calendric Association Rule [6], which is an optimization on "cyclic association rule" to capture real-life complicated temporal patterns. Nair et al. (2015) [7] also used support in their approach where they used Symbolic Aggregate approXimation (SAX)–Apriori based stock

* Corresponding Author

MileTS '19, August 5th, 2019, Anchorage, Alaska, USA

https://doi.org/10.1145/1234567890

trading recommender system to mine temporal association rules for stock price data. However, these approaches have not addressed adapting confidence in temporal mining. T-Apriori algorithm [4] is a modification of the Apriori algorithm, on transactional databases with the time constraint to generate rules for environmental systems.
Temporal association rule mining discovers rule within a given timeframe only. However, we want to see temporal relationships where an occurrence of one unusual event is linked to other unusual events happening simultaneously or after a certain period of time, i.e. a delayed effect. This motivated us to look into related works on delayed correlation. Yamtani et. Al. (2014) [7]used Delayed Correlation Analysis (DCA) to analyze the software evolution with the assumption that change in one variable during certain time period will affect other variables after some time delay. Liang et al. (2015) [8] used Generalized Cross Correlation (GCC) method on infrasound signal to estimate the time delay.

In our approach, we employ the concept of overlap and proximity to discover direct and time-delayed relation. We use anomalous clusters discovered in all domains to find these relations. If anomalous cluster of one domain is directly overlapping with the other, then we identify direct relations for them. If not we look for proximity between anomalous cluster sets and identify relation between those domains after shifting one domain by a certain time-delay width.

3 Methodology

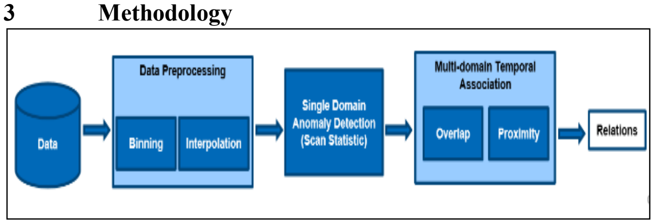

Figure 1: Multi-Domain Anomalous Temporal Association (Multi-DATA)

Figure 1 above presents the overall approach for Multi-DATA. The first step is data preprocessing in which we use binning and interpolation. Since data discretization segregates data into smaller sections, scan statistic can discover anomalies well because it can discover anomaly based on the normal/anomalous range for a smaller section rather than generalizing the range for the entire data. After binning, we discover anomalous windows in each bin for individual domains using scan statistic. We then look for temporal associations where we employ the concept of overlap and proximity and discover relations between multiple distinct domains. We next describe certain non-standard aspects of our approach in more details.

3.1 Anomaly detection

During single domain anomaly detection our goal is to capture points or subsequences of events that are not normal with respect to the others. We utilized temporal scan statistics for single domain anomaly detection (Kulldorff, 2001)[9]. We believe that these unusual series of events often contain interesting knowledge. Hence, we capture these anomalous windows from each of the domains being analyzed and mine the knowledge extracted from them for further analysis.

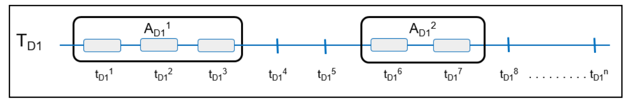

Figure 2: Anomalies in time series domain D1

In Figure 2 we can see that for a time-series domain $T_{D1} = \{t_{D1}^1, t_{D1}^2, \ldots\, t_{D1}^n\}$, where D1 represents the first time-series domain and t_{D1}^n is a time event recorded at time n, a set anomalous windows is represented as $A_{D1} = \{A_{D1}^1, A_{D1}^2, \ldots A_{D1}^i\}$, where $A_{D1}^i = \{t_{D1}^{n-p}, t_{D1}^{n-p+1}, \ldots\, t_{D1}^{n-q}\}$ is i^{th} anomalous window of the first domain and A_{D1}^i contains a subsequence of time events between t_{D1}^1 and t_{D1}^n.

3.2 Association of overlapped anomalous windows

Once the anomalous windows for each distinct domain are discovered, the next step is to discover and quantify relations between these domains using the anomalous windows. In our approach, we take a set of anomalous windows from all distinct time-series domains and use association rule mining to discover the relation between these domains. Before applying the algorithm, we first identify the number of overlaps, as explained in definition 1, between anomalous windows across domains. If more than a certain percentage of anomalous windows pairs overlap, then we discover associations across them. If not, we investigate delayed correlation in anomalous windows across domains.
DEFINITION 1: *[Overlap] Let t_x and t_y be time windows from domain x and y respectively. For time windows $t_x = \{t_x^1, \ldots t_x^n\}$ and $t_y = \{t_y^1, \ldots t_y^m\}$ overlap O_{xy}^i between t_x and t_y exists if both time windows have at least one identical time event i.e. $t_x^n = t_y^m$.*
Overlaps between anomalous time windows from two distinct domains mean some unusual activities happening in those domains during the same time period as shown in Figure 3. We assume that overlap indicates co-occurrence relation between these distinct domains. However, overlaps can also occur due to a coincidence. To avoid discovering such overlaps we set a threshold for the number of identical time events in an anomalous time windows pair and the number of bins with overlapping anomalous time windows. We also plan to perform Monte Carolo Simulations to eliminate the possibility of randomized occurrences. For a pair of anomalous time windows in a bin, from distinct domains, if more than 50% of total time events in each anomalous time windows are identical then they are

considered to have an overlap. If more than 50% of total number of bins have anomalous time windows pairs with overlaps, then a set of domains are considered to have significant overlaps.

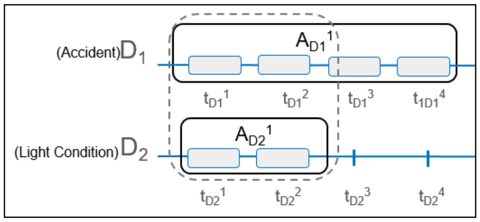

Figure 3: Anomalous time window overlap

DEFINITION 2: *[Proximity] For n number of bins, let us consider a pair of anomalous time windows with t_x and t_y, where t_x and t_y are anomalous time windows in the nth bin from domain x and y respectively. Let d_{xy} be the distance between t_x and t_y. Proximity P_{xy} is defined as the threshold used to determine the nearness between two time windows, t_x and t_y. It is calculated as $P_{xy} = T/(n*2)$, where T is the total number of time events in either domain, and $T = T_x = T_y$ and n is the number of bins. Time window t_y is said to be in proximity with respect to t_x if $P_{xy} > d_{xy}$.*

Time windows within proximity, as outlined in definition 2, are considered neighbors. If no overlaps or overlaps in less than half of anomalous windows pairs are found, then we check if those pairs are within proximity or not. Based on the existence of proximity, we check for the delayed relation for the set of domains.

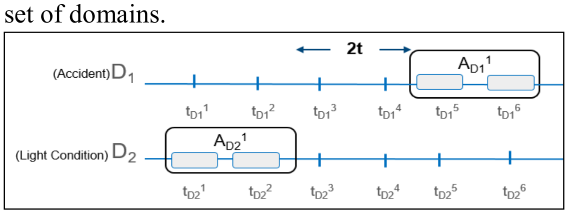

Figure 4: Proximity of 2t

As we can see in Figure 4, anomalous windows A_{D1}^1 is said to be within proximity with respect to A_{D2}^1 if proximity, $P \geq 2t$.

Figure 5 illustrates the technique of discovering multi-domain associations where A_{Di}^x represents anomalous windows and dotted lines represent overlaps. For the time-series domain, traffic, we have anomalous windows, $A_{D1}^1 = \{t_{D1}^1, t_{D1}^2, t_{D1}^3, t_{D1}^4\}$ and $A_{D1}^2 = \{t_{D1}^7, t_{D1}^8, t_{D1}^9\}$, where t_{D1}^n is an unusual time event recorded at time n. For another time series domain, air toxicity, we have anomalous windows, $A_{D2}^1 = \{t_{D2}^1, t_{D2}^2, t_{D2}^3\}$, $A_{D2}^2 = \{t_{D2}^6, t_{D2}^7, t_{D2}^8\}$ and $A_{D2}^3 = \{t_{D2}^{10}\}$. We can see that anomalous windows for these domains are overlapped at t^1, t^2, t^3, t^7, and t^8. Next, we generate a transaction where anomalous temporal events are treated as a transaction and domains with an anomaly in those temporal events are treated as items in a normal transaction. We then utilize the Apriori algorithm to compute the association, support, confidence and lift.

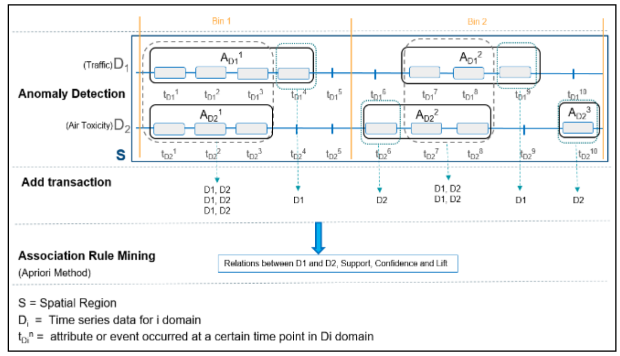

Figure 5: Multi-domain anomaly association framework

For each bin, we also check for delayed correlation between anomalous time windows if they are within certain proximity. We check if anomalous windows in a bin are correlated by using cross-correlation with lag of δ, then we identify the time lag with maximum correlation δ_{max} and shift a domain with the δ_{max} value. We then create transactions and use Apriori to discover associations.

4 Experimental results

We used two multi-domain real-world datasets MATCH (Mobilizing Action Toward Community Health) [11], NJDOT (New Jersey Department of Transportation)[12], and weather data[13] to experiment and validate our approach. We also used a synthetic data to allow us to measure the performance of our approach. Due to space constraints, we next discuss only a few key findings.

Multi-DATA associations: Figure 6 (a) shows associations in NJDOT data, across multiple bins, between Light and Surface Condition with confidence of 1 and lift of 5. We also observe significant overlaps between the domains

Time delayed associations: To find out if the set of domains have *time-delayed associations*, we check if anomalous window pairs of those domains are within proximity or not. If the set of anomaly pairs are within proximity, then we further analyze them to check for the delayed relation. In weather, data there are limited set of overlaps, as shown in figure 6 (b), so we further analyzed this data for time-delayed correlations. We computed the cross-correlation between each domain using the lag of δ. We used δ = 43 because we are using one data with 698 days and binning it into eight bins, which makes the size of time events about 86 in each bin so, we used half the size of bin for δ. Then we shift one domain by a width of time-delay constant δ_{max}, which is the lag with maximum correlation value. We also explored correlograms [10]. As shown in figure 7, x-axis gives the lag and y-axis gives the correlation, r_δ at each lag represented by vertical bars in the plot. Horizontal dotted line indicates confidence interval (CI), which is set to 90%. We found out that all anomaly pairs were within the defined proximity.

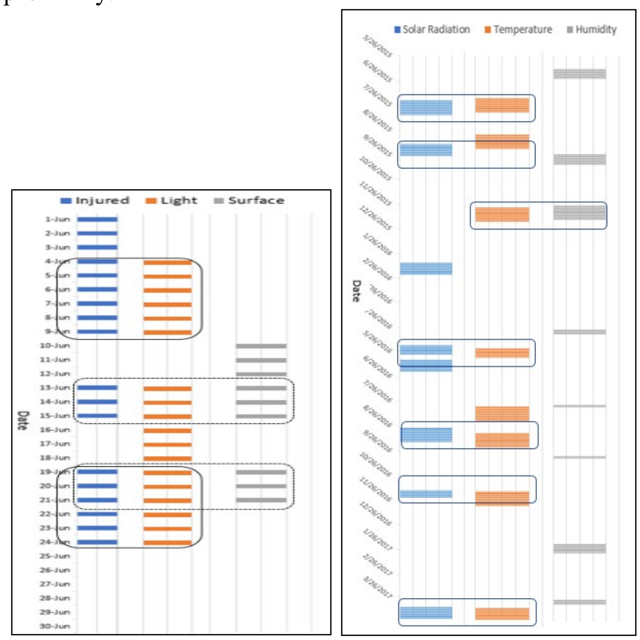

Figure 6: Anomalies in (a) NJDOT (b) weather data

From the Figure 7, we can see that highest correlation value is at lag 15, which is our δ_{max}. Hence, we shift one domain with time event width of 15t and discover the associations. This indicates that Humidity has delayed correlation with temperature. We also observed that Humidity has delayed relation with both temperature and solar radiation.

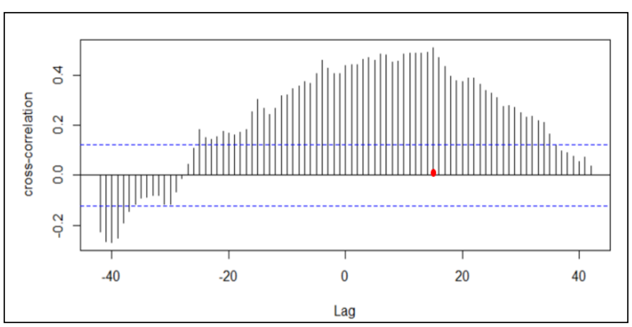

Figure 7: Correlogram between Temperature and Humidity in Bin 2

Single Domain Anomaly Detection: We performed comparison of single domain anomaly detection in synthetic data where we imputed anomalies. Our goal here is to measure how many of anomalous time events were discovered with scan statistics.

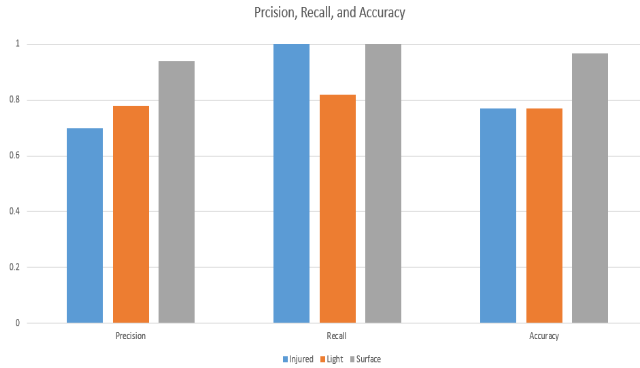

Figure 8: Single domain anomaly detection: Precision, Recall and Accuracy

We can see in figure 8 that scan statistics achieved a good result for Surface Condition. However, for Total Injured and Light Condition we see relatively lower values for Precision and Accuracy. Recall for all domains is high. Lower values of Precision and Accuracy may indicate that scan statistic captured False Positives in anomalous windows, which could be due to existing anomalous time units in real-world data where we imputed synthetic values. We also plan to explore other single domain anomaly detection methods to improve the accuracy of this first step, which can influence the results from the overall method.

Additional Validation: We performed piecewise aggregate approximation on each domain and found that the mean values were similarly high where overlaps were expected.

We compute the correlation between the anomaly pairs where associations are found. For NJDOT and synthetic data, we had overlaps in more than 50 % of anomalous window pairs, which implies stronger direct relation between domains in that dataset. Therefore, we computed correlation as another performance measure and found clear positive correlation between the anomaly pairs.

5 Conclusion and Future works

This paper proposed a novel algorithm to discover temporal associations across multiple distinct domains using time windows with unusual events. Our proposed algorithm allows to explore complex real-world linkages across domains. We employed the concepts of overlap and proximity to discover the direct or time-delayed relations across domains. In our future work, we plan to extend this approach to evaluate *n* temporal domains with different time-resolutions and present comparisons with relevant approaches in climate science where extreme value time series are evaluated.

REFERENCES

[1] Janeja, V. P., & Palanisamy, R. (2013). Multi-domain anomaly detection in spatial datasets. *Knowledge and information systems*, *36*(3), 749-788.

[2] Antunes, C. M., & Oliveira, A. L. (2001, August). Temporal data mining: An overview. In *KDD workshop on temporal data mining* (Vol. 1, p. 13).

[3] Roddick, J., Spiliopoulou, M.: A Survey of Temporal Knowledge Discovery Paradigms and Methods. In IEEE Transactions of Knowledge and Data Engineering, vol. 13, 2001.

[4] Liang, Z., Xinming, T., & Wenliang, J. (2005, August). Temporal association rule mining based on T-Apriori algorithm and its typical application. In *Intl. Symposium on Spatial-Temporal Modeling Analysis* (Vol. 5, No. 2).

[5] Harms, Sherri K, Temporal Association Rule Mining in Event Sequence, InEncyclopedia of Data Warehousing and Mining, ed. John Wang, 1098-1102 ,2005 Ramaswamy, S., Mahajan, S., & Silberschatz, A. (1998, August). On the discovery of interesting patterns in association rules. In *VLDB* (Vol. 98, pp. 368-379).

[6] Nair, B. B., Mohandas, V. P., Nayanar, N., Teja, E. S. R., Vigneshwari, S., & Teja, K. V. N. S. (2015). A Stock Trading Recommender System Based on Temporal Association Rule Mining. *SAGE Open*, *5*(2), 2158244015579941.

[7] Yamtani, Y., & Ohira, M. (2014, November). An Exploratory Analysis for Studying Software Evolution: Time-Delayed Correlation Analysis. In *2014 6th International Workshop on Empirical Software Engineering in Practice (IWESEP)* (pp. 13-18). IEEE.

[8] Liang, M., Xi-Hai, L., Wan-Gang, Z., & Dai-Zhi, L. (2015, September). The Generalized Cross-Correlation Method for Time Delay Estimation of Infrasound Signal. In *2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC)* (pp. 1320-1323). IEEE.

[9] Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *164*(1), 61-72.

[10] Metcalfe, A. V., & Cowpertwait, P. S. (2009). *Introductory time series with R*.

[11] "County Health Rankings." mobilizing action toward community health (MATCH) project

[12] New Jersey accident data for state routes: http://www.state.nj.us/transportation/refdata/accident/

[13] Beach Weather Stations - Automated Sensors | City of Chicago | Data Portal. (n.d.). Retrieved April 25, 2017, from https://data.cityofchicago.org/Parks-Recreation/Beach-Weather-Stations-Automated-Sensors/k7hf-8y75