Progression Analysis of Signals: Extending CRISP-DM to Stream Analytics

Pankush Kalgotra
MSIS Department
Oklahoma State University
Stillwater, OK-USA
pankush@okstate.edu

Ramesh Sharda
MSIS Department
Oklahoma State University
Stillwater, OK-USA
ramesh.sharda@okstate.edu

***Abstract*—Stream analytics focuses on analysis of signals generated simultaneously and over time. The specific patterns in the signals can indicate some of the outcomes such as failure of a device, etc. Therefore, novel ways to find specific patterns in the signals generated by many sources are required. In this paper, we extend the CRISP-DM process to include data preparation approaches for sequence mining. We present progression analysis, an approach for converting streams of records to be able to detect useful signals for analysis. To illustrate the process, we present a healthcare example where patients diagnosed with Tobacco Use Disorder develop multiple other diseases over multiple hospital visits. The common sequences of the diseases diagnosed in the TUD patients over multiple hospital visits are presented and discussed. Finally, the generalizability of the progression analysis is discussed.**

Keywords—multidimensional time variant; streaming analytics; tobacco use disorder; sequence analysis

I. Introduction

Several Big Data sources such as the web, sensor technologies, mobile technologies, etc. generate variety of data streams at high velocity. Storing, securing and analyzing the data streams has become one of the most critical issues for data scientists. To store and secure the huge volume of variety of data streams generated with high velocity, efficient filtering and security measures are necessary [1]. Moreover, Big Data tools and technologies are required to analyze massive data from variety of streams [2].

Stream analytics is an interesting and challenging subset of Big Data problems. Streams are used in different domains and thus, it is a useful area to explore with respect to Big Data process methodology. To do a streaming analytics project effectively and efficiently, a general process or methodology is necessary [3]. We adopt the traditional CRoss-Industry Standard Process for Data Mining (CRISP-DM) [4] to analyze the time-stamped data streams. CRISP-DM involves six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The adoption of CRISP-DM to analyze streams will extend its applicability to multiple other domains.

All six phases of CRISP-DM are applicable to the streaming analytics. However, data preparation and modeling phases of the streaming data are different from the traditional static data mining due to its time-series nature. Researchers in the past have largely focused on proposing the models for the data, which is important. But at the same time, novel approaches for preparing the time variant data are also required to organize data for modeling. As the data in raw form is difficult to analyze, it is crucial to convert the raw data in the form of a structured dataset to be able to perform any kind of analysis.

Due to the time variant multidimensional nature of the streams, the traditional data mining techniques might not be capable of analyzing them. The novel temporal methods based on the evolutionary algorithms are required [5]. The methods should be able to synchronize multidimensional data streams and generate meaningful patterns. In this paper, we present an algorithm to prepare multidimensional data streams in a form to perform sequence analysis. It is an efficient way to analyze time-ordered data [6, 7]. The resultant data format can allow data mining approaches to find sequential patterns in the time-ordered transactional dataset. Different patterns of sequences of events within the transactions are extracted based on some user-specified minimum support (support is the percentage of sequences that contain the specific pattern).

The multidimensional data streams are very common in different systems. One example of such systems belongs to the healthcare sector. When multiple diseases are diagnosed in a patient over multiple hospital visits, the occurrence of multiple diseases over time make the data multidimensional and time-variant (MDTV). Later in this paper, we illustrate the proposed method on the time-ordered dataset containing the patients diagnosed with Tobacco Use Disorder.

Another area of application of multidimensional time-order streams is Internet of Things (IoT). The elements/devices/sensors in IoT produce variety of data streams at high velocity. The components of an IoT platform share information by exchanging signals. Devices may send multiple signals simultaneously over time. Hence, the scenario is an MDTV.

Next, we describe a hypothetical problem to describe the multidimensional time-order streams followed by the method proposed to prepare the data for analysis. The method is illustrated using the Electronic Medical Records of the patients with Tobacco Use Disorder. Finally, the results are discussed followed by the concluding remarks.

II. Scenario

We present a specific scenario in which a device or any element of a system emits multiple signals over time. Moreover, it sends multiple signals simultaneously. A hypothetical system is presented in Figure 1 in which a device S is sending signals to a device D. The device S is sending the signals A and B simultaneously, followed by C and D, finally E and F. As more devices are added to both input and output sides in Figure 1, the system gets complex and it becomes difficult to understand the sequences of signals. By developing a process to prepare the streams and finding the temporal patterns in the MDTV data, we contribute to the method of analyzing the data streams.

Fig. 1. A Hypothetical IoT System

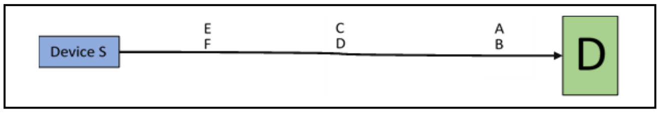


TABLE I. Dataset corresponding to Figure 1

From	To	Time	Signal
S	D	1	A
S	D	1	B
S	D	2	C
S	D	2	D
S	D	3	E
S	D	3	F

III. MDTV Data Preparation Method for Sequence Analysis

To begin with, the data need to be converted into semi-structured or structured form. The streaming data generated from the system in Figure 1 need to be cleaned and structured as presented in Table 1. It includes the data-generating device, time of the signal and signal itself.

As our aim is to prepare data for sequence mining, the streaming data is required to be converted into a transactional dataset. To do so, the time-ordered dataset is separated into different buckets. This process is called sessionizing the events or signals. The concept of a session is highly used in web analytics to connect a series of user events online. A session is defined in multiple ways by the researchers in web analytics [8, 9]. To define a few, a sequence of events or queries are considered as one session if 1) these occur within specific time, for example, one hour; 2) no more than t time passes between successive events. This t time is known as the "period of inactivity" or the "timeout threshold"; and 3) the collections of signals to complete a particular task are considered. In the MDTV case, any of the following approach can be followed to sessionize the signals generated by the signals.

The steps followed to prepare the data for finding sequential patterns in the signals is presented in Table 2. The process is illustrated with an example in Table 3a to Table 3d. To illustrate the process, we use a hypothetical dataset containing a device, S, emitting signals at three occasions in Table 3a. At time 1, two events (A and B) are emitted simultaneously; at time 2, two new unique events are generated (C and D); and at time 3, E and F are emitted simultaneously. Our aim is to create the sequence of signals as shown in Figure 2.

To create a dataset appropriate for generating sequences of events as in Table 3d, we first divided the dataset into three datasets, each having information about a time as shown in Table 3b. The left join of time 1 dataset with the time 2 dataset is performed, and its output is further left joined time 3 dataset. In our hypothetical dataset, we have three time slots. However, one can extend the analysis to any number. The resultant table from multiple joins is presented in Table 3c, in which each combination of the events is labeled by a different session. Finally, the dataset is converted as in Table 3d by appending information on signals at different time intervals. This dataset is suitable for running sequence analysis.

The dataset prepared in Table 3d is taken as input by the sequence analysis. The sessions in Table 3d are analogous to the transactions in the sequence mining method. Finally, sequence analysis can be run to discover common sequence patterns.

To visually study the common sequences, the nPath function of Teradata Aster, a Big Data platform can be used. This function is able to create paths with a session ordered by the time. However, the count of a path is adjusted based on its

TABLE II. Steps to find prepare data for sequence analysis

Step 1: For n data sources (i=1 to n) with v times and k signals each time, the dataset contains
P_i, $V_{ij\ (j=1\ to\ v)}$, $D_{ijk\ (k=1\ to\ m)}$

Step 2: A separate dataset for each time is extracted; a total of v number of datasets are resulted:
(P_i, $V_{ij\ (j=1)}$, $D_{ijk\ (k=1\ to\ m)}$), (P_i, $V_{ij\ (j=2)}$, $D_{ijk\ (k=1\ to\ m)}$) … P_i, $V_{ij\ (j=v)}$, $D_{ijk\ (k=1\ to\ m)}$

Step 3: The dataset with J=1 is left joined with J+1. The resultant dataset contains
(P_i, (V_{ij}, $D_{ijk\ (k=1\ to\ m)}$ $)_{(j=1)}$, (V_{ij}, $D_{ijk\ (k=1\ to\ m)}$$)_{(j=2)}$… ($V_{ij}$, $D_{ijk\ (k=1\ to\ m)}$$)_{(j=v)}$). *In each join,* $D_{jk} \neq D_{(j-l)k}$, *where l=1 to (j-1).*

Step 4: Each record is assigned a unique session. Total number of sessions per data source
$S=\prod_1^v C_i$, *where C is count of signals at a time i.*

Step 5: A union of P_i, V_j, D_K *and* S_S *are taken as*
(P_i, $V_{ij\ (j=1)}$, $D_{ijk\ (k=1\ to\ m),}$ $S_{(1\ to\ s)}$) U (P_i, $V_{ij\ (j=2)}$, $D_{ijk\ (k=1\ to\ m),}$ $S_{(1\ to\ s)}$) U … (P_i, $V_{ij\ (j=v)}$, $D_{ijk\ (k=1\ to\ m),}$ $S_{(1\ to\ s)}$)

Step 6: The Sequence Analysis is run to find the signal progressions.
Step 7: The number of sequences are adjusted by calculating the number of data sources following the specific signals sequences from the result of Step 3.

Device	Time	Signal
S	1	A
S	1	B
S	2	C
S	2	D
S	3	E
S	3	F

TABLE IIIa

Device	Time	Signal
S	1	A
S	1	B

Device	Time	Signal
S	2	C
S	2	D

Device	Time	Signal
S	3	E
S	3	F

TABLE IIIb

Device	Time_1	Signal_1	Time_2	Signal_2	Time_3	Signal_3	Session
S	1	A	2	C	3	E	S1
S	1	B	2	C	3	E	S2
S	1	A	2	D	3	E	S3
S	1	B	2	D	3	E	S4
S	1	A	2	C	3	F	S5
S	1	B	2	C	3	F	S6
S	1	A	2	D	3	F	S7
S	1	B	2	D	3	F	S8

TABLE IIIc

Device	Time	Signal	Session
S	1	A	S1
S	2	C	S1
S	3	E	S1
S	1	B	S2
S	2	C	S2
S	3	E	S2
S	1	A	S3
S	2	D	S3
S	3	E	S3
S	1	B	S4
S	2	D	S4
S	3	E	S4
S	1	A	S5
S	2	C	S5
S	3	F	S5
S	1	B	S6
S	2	C	S6
S	3	F	S6
S	1	A	S7
S	2	D	S7
S	3	F	S7
S	1	B	S8
S	2	D	S8
S	3	F	S8

TABLE IIId

Adjusted nPath

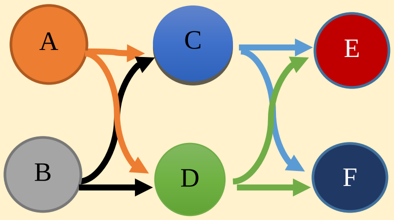


Fig 2. Sequence patterns

Progression Analysis Process

original numbers. For instance, there is only one path between A and C, but according to the Table 3d, there are two paths between A and C.

IV. Illustration

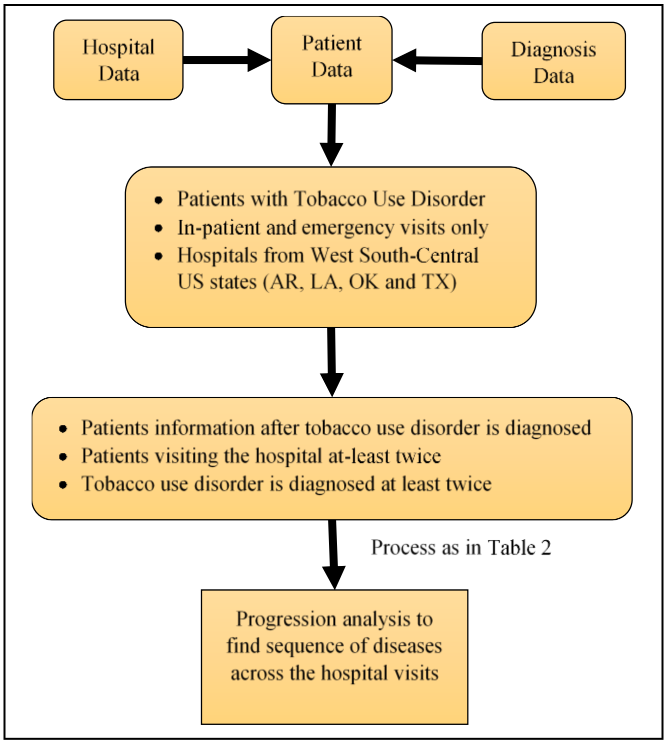
Our method can be applied to find sequences in the events generated by any entity generating simultaneously signals over time. To illustrate our method, we present a healthcare application. Specifically, we present the sequences of diseases over hospital visits developed by the patients. This particular example is analogous to the system described in the previous section. Like a device generates multiple signals over time, a patient also develops multiple diseases over hospital visits. One or a cluster of diseases may lead to another disease, which is diagnosed in the next hospital visit.

For the illustration, we focus on the patients from South-Central regions of United States (AR, LA, OK and TX) who are diagnosed with Tobacco Use Disorder. We obtained the dataset from the Cerner Corporation, a major Electronic Medical Records (EMR) provider. The database is housed at Center for Health Systems Innovation (CHSI) in Oklahoma State University. The database contains EMRs of the visits of more than 50 million unique patients across US hospitals (1999-2014). In this paper, we focused on EMRs of the patients diagnosed with the Tobacco Use Disorder.

A. *Data Preparation*

First of all, the information about the patients, hospitals, types of visits and diseases developed by the patients were combined for further analyses. Next, the emergency and inpatients diagnosed with the Tobacco Use Disorder were extracted to find the comorbidities. As TUD is developed at a certain point in the life of a patient, we considered his/her hospital visits from the first time occurrence of the TUD onwards. It was also made sure that a patient visited the hospital at least twice and TUD was diagnosed at least in two visits. These measures were taken to reduce the sample bias in the data as TUD may not be a prevalent disease in a patient. If TUD is diagnosed at least twice, the disease sequences may be the indicative evidence of the TUD effect. After cleaning and preparation, the final dataset contained information about 25,330 patients. In the dataset, 56% of the patients were females; 75% were Caucasians, 19% African-American, 2.5 % Hispanics, 2.5% Native Americans and less that 0.2% were Asians. The step-by-step process followed to prepare and analyze the data is presented in Figure 3.

Fig. 3. Data Preparation and Analysis Step


Hospital Data
Patient Data
Diagnosis Data
Patients with Tobacco Use Disorder
In-patient and emergency visits only
Hospitals from West South-Central US states (AR, LA, OK and TX)
Patients information after tobacco use disorder is diagnosed
Patients visiting the hospital at-least twice
Tobacco use disorder is diagnosed at least twice
Process as in Table 2
Progression analysis to find sequence of diseases across the hospital visits

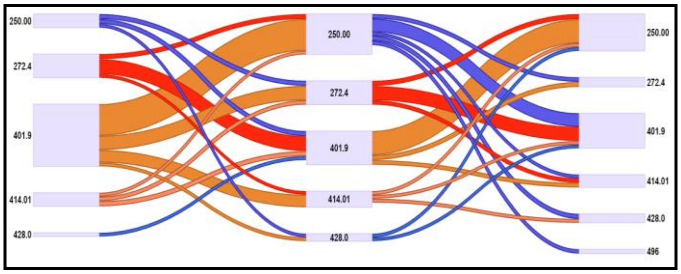
V. Results

To illustrate the process, we restrict the sequence to the third visit in the hospital. So, maximum number of diseases in a sequence will be three. Moreover, we present top twenty sequences to understand the paths between the diseases over three visits. The visualization of sequence of diseases is presented in Figure 4 and paths with percentage of patients following a specific sequence is listed in Table 3. Figure 4 is from left to right representing the three visits at each box. The width of an arc between two disease represents its frequency of occurrence.

Figure 4 presents Hypertension (ICD9 – 401.9) as the most prevalent diseases in TUD patients on the first hospital visit. It is followed by Diabetes mellitus (ICD9-250.00) on the second visit by more than 45% patients; Hyperlipidemia by about 24% patients, Coronary atherosclerosis by 21% patients and about 10% patients developed Congestive heart failure during second visit after hypertension on the first visit. Similarly, other sequences in Figure 4 can be interpreted.

Across the three visits, the most common sequence is Hypertension (ICD9-401.9), followed by Hyperlipidemia (ICD9-272.4) on the second visit and Diabetes mellitus (ICD9-250.00) on the third visit. (This can be observed in the first row of Table 4.) About 16% of the TUD patients followed this specific path. The second most sequence in Table 4 is Hyperlipidemia to Hypertension to Diabetes mellitus. The third

Fig. 4. Top 20 sequences of diseases over three visits



common sequence is Diabetes mellitus to Hyperlipidemia to Hypertension. These sequences show that Hyperlipidemia, Hypertension and Diabetes mellitus are highly related to each other across visits. Similarly, other sequences can be interpreted. As multiple diseases can be present at the same time, one patient can exhibit multiple sequences and therefore the sum of percentages is more than 100.

TABLE IV. TOP 20 SEQUENCES OF DISEASES OVER THREE VISITS

S. No.	Sequence	Percent
1	[401.9, 272.4, 250.00]	16.27
2	[272.4, 401.9, 250.00]	15.49
3	[250.00, 272.4, 401.9]	15.47
4	[401.9, 250.00, 496]	14.65
5	[414.01, 401.9, 250.00]	14.63
6	[272.4, 401.9, 414.01]	14.24
7	[401.9, 250.00, 428.0]	14.24
8	[401.9, 272.4, 414.01]	14.12
9	[250.00, 401.9, 272.4]	13.96
10	[414.01, 401.9, 272.4]	13.95
11	[401.9, 250.00, 272.4]	13.86
12	[401.9, 414.01, 272.4]	13.75
13	[272.4, 250.00, 401.9]	13.73
14	[414.01, 272.4, 401.9]	13.64
15	[401.9, 414.01, 428.0]	13.64
16	[250.00, 401.9, 428.0]	13.35
17	[401.9, 414.01, 250.00]	13.31
18	[272.4, 414.01, 401.9]	13.03
19	[428.0, 401.9, 250.00]	13.03
20	[401.9, 272.4, 428.0]	12.94

ICD-9 codes	Disease
272.4	Hyperlipidemia
401.9	Hypertension
250.00	Diabetes mellitus
414.01	Coronary atherosclerosis
428.0	Congestive heart failure
496	Chronic airway obstruction

VI. DISCUSSION AND CONCLUDING REMARKS

The above example is an illustration of the method of preparation presented in this paper. We showed the common sequences of different diseases diagnosed by the TUD patients over multiple visits. The sequences of diseases discovered can provide additional understanding about how different diseases develop sequentially in TUD patients. The knowledge about sequence of diseases can help physicians take preemptive actions to prevent the future diseases. The progression analysis can be helpful to create an expert system for predicting future diseases.

We restricted our analysis to finding paths. However, it will be more interesting to see the outcome of a specific path. For instance, it is possible that some specific sequences of diseases can lead to death of a patient but not others. Classifying such sequences will be part of our future research. In this study, we primarily focused on the data preparation and analysis phases of CRISP-DM. Our future work will include evaluation of our method versus the existing methods and the deployment our method in the real-world setting.

The application of progression analysis process is not limited to the healthcare problems. It is a generalizable process and can be applied to any situation in which multiple signals are generated over time. As discussed earlier, finding patterns in the sequences of the signals generated by the IoT devices can be insightful. For instance, it is possible that a device follows a specific pattern just before the failure of a device. An unexpected pattern of signals may forecast the future state of the IoT systems. If such patterns are known and detected timely, preemptive maintenance actions can be taken.

However, it has to be noted that the interval between visits in our healthcare example could be very large. In contrast, the interval between signals generated by the devices in the IoT environment is mostly very short. Hence, to apply progression analysis on fast generated signals, powerful Big Data tools are required. In our particular case, we used Teradata Aster, a Big Data Platform. In contrast with IoT environment, where all signals that are emitted are tracked by the system, only some of the patients' signals get noticed - only those that occurred during his/her visit. Between two visits the patient could have a disease that disappeared by the second visit.

Our contribution is in preparing the data to be able to perform modeling and analysis. Due to the growing streaming data in almost every discipline, generalizable methods such as presented in this paper can help enhance the power of data analytics.

ACKNOWLEDGMENT

We are thankful to Cerner Corporation, which has shared the electronic medical records data through the Center for Health Systems Innovation to enable us to carry out this analysis.

REFERENCES

[1]. Chui, M., M. Löffler, and R. Roberts, *The internet of things.* McKinsey Quarterly, 2010. **2**(2010): p. 1-9.

[2]. Strohbach, M., et al., *Towards a big data analytics framework for IoT and smart city applications*, in *Modeling and Processing for Next-Generation Big-Data Technologies*. 2015, Springer. p. 257-282.

[3]. Saltz, J.S. *The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness*. in *Big Data (Big Data), 2015 IEEE International Conference on*. 2015. IEEE.

[4]. Shearer, C., *The CRISP-DM model: the new blueprint for data mining.* Journal of data warehousing, 2000. **5**(4): p. 13-22.

[5]. Gubbi, J., et al., *Internet of Things (IoT): A vision, architectural elements, and future directions.* Future Generation Computer Systems, 2013. **29**(7): p. 1645-1660.

[6]. Agrawal, R. and R. Srikant. *Mining sequential patterns*. in *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*. 1995. IEEE.

[7]. Srikant, R. and R. Agrawal. *Mining sequential patterns: Generalizations and performance improvements*. in *International Conference on Extending Database Technology*. 1996. Springer.

[8]. Gayo-Avello, D., *A survey on session detection methods in query logs and a proposal for future evaluation.* Information Sciences, 2009. **179**(12): p. 1822-1843.

[9]. Spiliopoulou, M., et al., *A framework for the evaluation of session reconstruction heuristics in web-usage analysis.* Informs journal on computing, 2003. **15**(2): p. 171-190.