Seattle
Making Value for Society
122nd ASEE Annual
Conference & Exposition
June 14 - 17, 2015
Seattle, WA
Paper ID #13708

Application of Sequence Data Mining for Adverse Event Prediction and Action Recommendation

Dr. Reza Sanati-Mehrizy, Utah Valley University

Reza Sanati-Mehrizy is a professor of Computer Science Department at Utah Valley University, Orem, Utah. He received his M.S. and Ph.D. in Computer Science from the University of Oklahoma, Norman, Oklahoma. His research focuses on diverse areas such as: Database Design, Data Structures, Artificial Intelligence, Robotics, Computer Aided Manufacturing, Data Mining, Data Warehousing, and Machine Learning.

Dr. Ali Sanati-Mehrizy

Dr. Ali Sanati-Mehrizy is a Pediatric resident physician at Rutgers University - New Jersey Medical School in Newark, NJ. He is a graduate of the Milton S. Hershey Pennsylvania State University College of Medicine. He completed his undergraduate studies in Biology from the University of Utah. His research interests are varied and involve pediatric hematology and oncology as well as higher education curricula, both with universities and medical schools.

Paymon Sanati-Mehrizy, Icahn School of Medicine at Mount Sinai

Paymon Sanati-Mehrizy is currently a medical student at the Icahn School of Medicine at Mount Sinai. He completed his undergraduate studies in Biology from the University of Pennsylvania in May 2012. Currently, his research interests consist of higher education curricula, both with universities and medical schools.

Dr. Afsaneh Minaie, Utah Valley University

Afsaneh Minaie is a professor of Computer Engineering at Utah Valley University. She received her B.S., M.S., and Ph.D. all in Electrical Engineering from University of Oklahoma. Her research interests include gender issues in the academic sciences and engineering fields, Embedded Systems Design, Mobile Computing, Wireless Sensor Networks, Nanotechnology, Data Mining and Databases.

Application of Sequence Data Mining for Adverse Event Detection and Action Recommendation

Abstract

Many real-life data mining applications use sequence data modeling, in which data is represented as a sequence. A temporal sequence is a finite ordered list of events (t_1,e_1), (t_2,e_2), …,(t_n,e_n) where t_i represents time and e_i represents the event taking place at time t_i. e_i takes place before e_{i+1} for $1\leq i \leq n-1$. This model can be used in data mining, called sequence data mining, to predict certain event that may take place at a specific time. Sequence data mining has a wide range of applications. This data mining technique can be used for prediction of adverse events and can recommend appropriate actions to be taken as needed. In aviation safety, the future of a flight can be predicted as a sequence and proper action can be recommended to avoid dangerous situations that a flight may get into otherwise. In health care, the future of a bacterial infection can be predicted and proper medicine can be prescribed for different situations to bring the patient's illness to an end. In marketing, customer shopping can be monitored and certain actions can be taken, such as mailing coupons, to encourage customers to engage in repeat shopping. In manufacturing, sensor data can be analyzed to regulate operations and predict and avoid dangerous situations by recommending appropriate actions.

This paper which is the continuation of the work by Sanati et. al.[1], discusses sequence representation, implementation, and its application for a number of different fileds.

Introduction

Data mining combines tools from statistics and artificial intelligence (such as neural networks and machine learning) with database management to analyze large digital collections, known as data sets[2]. It is a well-researched area of computer science with high demand due to its usefulness in any field with large quantities of data, where meaningful patterns and rules can be extracted.[3]. Therefore, many organizations and businesses can benefit from data mining techniques, as these organizations record lots of data daily.

Sequence data mining is a specific area of data mining that has a wide range of application in this field[4,5]. In some application, sequence data mining can be used to identify anomalous events and recommend proper actions to be taken to avoid such situations. In aviation safety, for example, the future of a flight can be predicted as a sequence of events. If a sequence of events appears anomalous, proper action(s) can be recommended to avoid dangerous situations that this flight may otherwise get into. In health care, physicians can predict the future of a bacterial infection or an allergic reaction. In a desire to bring the patient's illness to an end, physicians prescribe proper medication considering the situations. In other real-life situations such as operating manufacturing factories, sensor data can be analyzed to control operations, predict dangerous situations, and recommend and implement proper actions. This type of data analysis can help engineers estimate the remaining life of equipment and recommend proper maintenance services before the equipment malfunctions, eliminating costly delays.

Definition

A sequence is a nonempty ordered list of tuples (t_1, $e_{1,}$ a_1)…(t_n, $e_{n,}$ a_n) where t_i represents a time point, e_i represents an event and a_i represents an action at that time point. If the event e_{i+1} exists, it is the effect of event e_i. At any time point t_i, the event e_i may cause a set of possibly empty event(s). Any of these new events can initiate another sequence. A sequence can come to end if no new event is generated. In some applications, it is beneficial to bring a sequence to a halt situation. But in some other cases, it is desirable that an event causes another event that may initiate another sequence. At any time t_i a sequence may be brought to a halt situation if a proper action a_i is taken or it may come to an end without the need for an action taken by an outsider. In this case a_i is null.[1]

Classification of Sequences

Han et al.[6] have given a good coverage of sequence mining, including classification of sequences. The classification depends on what criteria are used for classification. In their classification, sequences are considered as complex data type.

Sanati[1] simply classifies sequences as desirable sequences and undesirable sequences just for the purpose of his research work. Desirable sequences are those that their existence is beneficial and one would want them to exist. Undesirable sequences are those that their existence is harmful and they need to be brought to an end as soon as possible. In marketing, customers' shopping behavior can be monitored and certain action can be taken, such as mailing coupons, to encourage the customer to continue shopping of relevant products that actually may initiate new sequence(s). A radio station may follow the listening habits of clients of different age groups and reward the groups by adjusting the broadcasting line-ups and infuse appropriate advertisements in order to boost profits for corporations. In these cases, the sequence is enhanced and augmented for expected positive outcomes. In health care, physicians can predict the future of a bacterial infection or an allergic reaction. These types of sequences are harmful and need to be brought to an end as soon as possible. To do so, physicians recommend necessary treatments to terminate these undesirable sequences.

Implementation Model

A multiway lexicographic search tree can be used to represent event sequences where an event from the sequence of events determines a multiway branch at each step. If the sequence is constructed from the English alphabets, at the root of the tree there are 27 possible branches. Similarly, there are 27 braches for each subsequent node of the tree. For the sake of simplicity, assume we have a text that its words are constructed from the letters a, b, and c. The multiway tree shown in Fig.1[7] represents sequences constructed from these letters. This tree can recognizes all three letter words as long as the letters are a, b, and c.

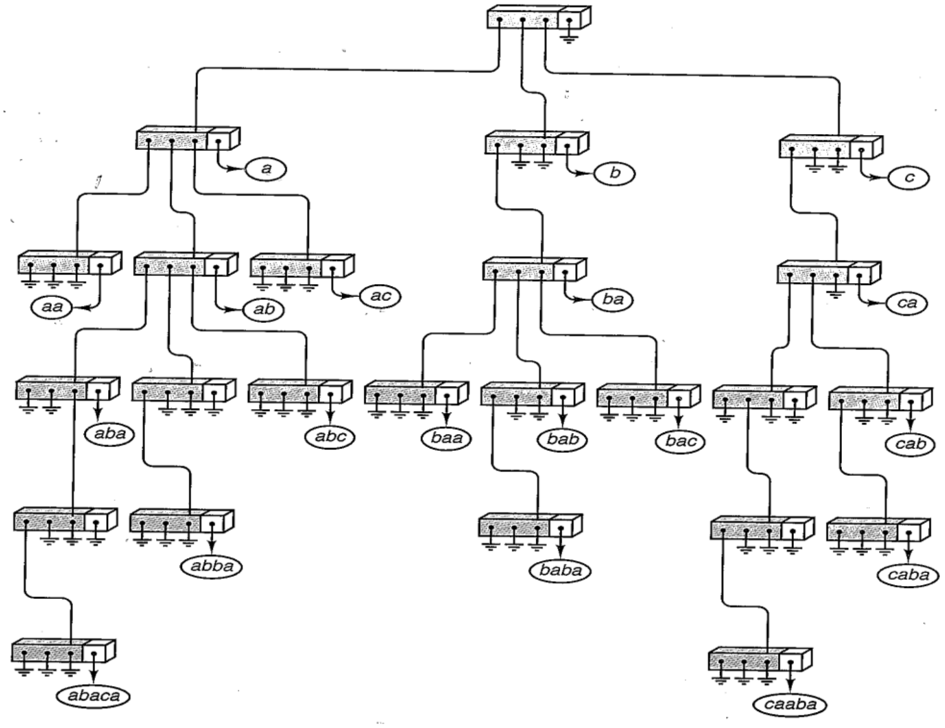

Fig. 1: A multiway lexicographic search tree constructed using the English letters a, b, and c.

At the root of the tree, there are four branches, one for each of the letters a, b, and c and a null pointer. At the next level, there are three nodes, and each node represents three branches for further options and one branch to show the sequence up to this point. At any node, the sequence is a concatenation of branch labels along the path from the root to this node.

For information retrieval, where these sequences represent keys of records, each node contains a pointer to the record of information with this sequence as its key. This data structure can be used to represent the sequences of events, where each branch of a node represents an event and each path in the tree represents a sequence of events. When a path terminates, it means that the corresponding sequence of events came to an end. Each node of the tree can be considered the outcome of a sequence at a particular time. The concatenation of events along the path from the root of the tree to this node is the cause of this outcome. At any node of the tree, the outcome can be evaluated. If the outcome is undesirable, we can look back for the cause of it and tune the system for a better outcome. If the label for each branch is a word, this model can recognize phrases and even sentences.

Some Applications of this Model

For the purpose of adverse event detection, where a sequence is constructed from events, nodes contain more branches, and each node contains a pointer to an action (possibly a set of actions) that needs to be taken place for the event leading to that node. This action may prevent initiation of any further event or may promote generation of new sequences.

Marketing: In the following diagram shown in Fig. 2, a customer has purchased baby food. Coupons for baby diapers can be sent to him to encourage the customer to purchase diapers. Later on, when the baby gets older, coupon for toys and a bicycle can be sent to him, and he again may buy toys and a bicycle. Each of these events is a starting point of a sequence. Having such sequence presented, when a customer purchases baby food, we can predict that he will purchase baby diapers, toys and a bicycle at some point in time.

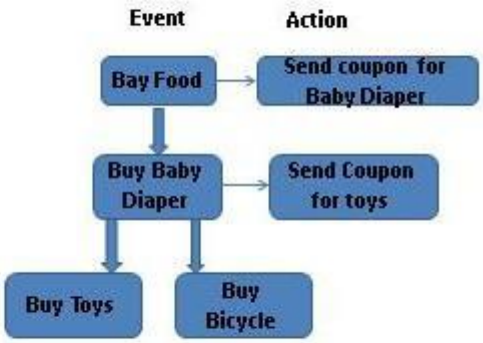

Fig. 2: The sequence of events and action recommendations in a customer shopping monitoring system

Health Care System: In terms of health care, sequence data mining can be useful for everything spanning patient management to operating a hospital network. Let us consider a patient showing signs of type I hypersensitivity (allergic) reaction due to inhalation of an allergen. The person is having frequent sneezes; soon the person may develop an itchy throat followed by runny nose and watery eyes, followed by asthmatic reaction.. If left untreated for more than a day, the patient may suffer delayed reactions of allergy that involves inflammatory reactions leading to itchy throat and skin, damage in the trachea and the alveoli of the lung. In a few days the person may get secondary bacterial infections of the lung that may lead to bronchiolitis, followed by high fever followed by pneumonitis and pneumonia. If left untreated, the person may develop septicemia, followed by vascular shock, followed by coma and ultimately death. This is an ordered sequence where each of these symptoms appears after a certain period of time. This is an undesirable sequence and needs to be brought to a halt situation as soon as possible. For example, interrupting the sequence by treating the subject with epinephrine when the patient developed runny nose and watery eyes could have stopped the sequence[8]. Had the sequence progressed beyond that step, specific actions could have been taken at the subsequent steps. For example, the patients could be treated with anti-inflammatory drugs within a day of the beginning of the episode or treated with antibiotic if secondary bacterial infection is diagnosed. If the recommended action is effective, it will cure the patient and terminates the sequence. If not, the progress continues and goes to the next step. In this situation, the prediction is easier because we can tell what the next step can be. For example, viral antigens may mutate, initiating a new sequence, making it more difficult to make accurate predictions.

Education: In some cases, we recommend an action to change the sequence intentionally. Consider a case when a student fails the first quiz and misses submitting the first project. The instructor can predict that this student will more likely have problems with other assignments, quizzes, exams and so on. The instructor may even guess that this student will be one of those who may fail the course. This is an example of an undesirable sequence. The instructor may take preemptive actions, such as contacting the student and trying to find out what the problem might be to make appropriate recommendations. If the recommendation is effective and the students changes his behavior, the sequence changes and hopefully the student may enter in more desirable sequence leading to successful completion of the course.

Adverse Event Detection

Based on previous sequences, a model can be designed for detecting future anomalous sequences. Nikunj C. Oza[9] presented a paper that includes a simple model for anomalous sequence detection. Consider the following sequences:

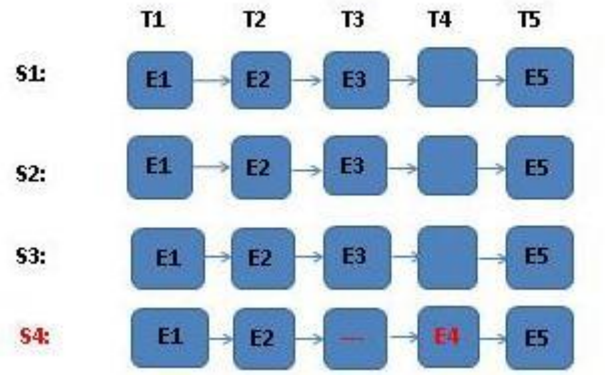

Fig. 3. Sequences

S1, S2, and S3 are normal sequences, but S4 is an anomalous sequence. Because it is expected that the event E3 take place at time t3 which did not happen and event E4 took place at time t4, which was not expected. This model can be used to detect anomalous flights in aviation safety system where events can be on/off positions of various switches.

Examples of applications of Sequence mining

Dong and Pei[11] have discussed some application of event sequence mining that we will present here.

Sequence Mining in Marketing

In a company, the marketing manager is responsible to increase the company’s market share and customer retention. To achieve these goals, the marketing manager wants to design a marketing campaign, which consists of two major tasks**.** First, a set of products should be identified for promotion. Hopefully, for promoting those products, customers will be retained, and sales on other products will be stimulated. Second, a set of customers should be targeted for marketing and promotion information delivery.

The manager has a list of transactions from the past. Each transaction includes the customer_Id, the list of products bought in the transaction, and the date and time of the transaction. Grouping the transactions by customer_Id and sorting them by the timestamp in ascending order, the manager can get a purchase sequence database where each sequence records the purchase behavior of a customer.

The manager may want to find frequent subsequences that are shared by many customers. As patterns, those frequent subsequences can help him to understand the shopping behavior of the customers and the associations among the products. Using these discovered purchase patterns and the target customers, the manager can identify the potential products to be promoted.

Sequence Mining in Aviation System

The safety manager in an airline is in charge of the braking system in airplanes. A sequence of status records is kept for each aircraft. Maintaining the braking system of an airplane in a hub airport of the airline is highly preferable because when the job is done in a guest airport, the service cost is often several times higher. On the other hand, being too proactive in maintenance may end up with unnecessary cost, since parts may be replaced too early and are not fully used.

Therefore, the manager is facing such a question: given an airplane's sequence of status records, predict in high confidence whether the plane needs maintenance before it goes to the next hub airport. This is a classification (a supervised learning) problem, since the prediction is made based on some historical data, that is, records of previous maintenance collected for references.

Sequence Mining in Healthcare System

A medical analyst in charge of analyzing patients' reactions to a new drug collects the sequence of reactions of the patients such as the changes in temperature, blood pressure, muscle pain, and so on. To be precise, data will be collected from many patients. So, there are a good number of such test cases, normally several hundreds or even thousands. In order to summarize the results, the analyst needs to categorize the cases into a few groups such that all cases in a group are very similar to each other and the case in different groups are substantially different from each other.

This is a clustering task known as unsupervised learning, because the sequences are not labeled in advance, and the groups should be defined by the analyst base on the similarity among the sequences.

Teaching Sequence Mining

In our computer science department, we teach an undergraduate data mining course. The major parts of workload for this class are exercise problems, data analysis projects, and research projects. Exercise problems and data analysis projects are individual works. For data analysis projects the students use Weka[10], a popular open source suite of machine learning software for data mining. But the research projects are team works usually with two team members on each team. Each team is given a certain topic in data mining to do research on. Each team has to present the result of its research to the class at the end of semester. This way, the teams share their work with others and learn from each other. Also, the students learn how to give a presentation in front of a good audience and answer questions. Naturally, this type of workload prepares student better for this highly competitive job market.

Finally, the teams will be encouraged to convert the result of their research project to a paper and submit it to a conference for presentation and publication. One problem that I feel I need to mention is the travel expense for the student to the conference. To solve this problem, usually the students can sign up for a few hours of volunteer work at the conference and get their registration fee waved. Sometimes, the students are willing to pay for a big part of the travel expenses because this trip becomes a vacation for them while securing a publication for their resume.

In order to enrich the content of this data mining course, the plan is for two weeks of teaching sequence mining be added to the content of this course. This gives the student some experience with mining such a complex data type that has applications in almost any field. In these two weeks the following topics will be covered:

Introduction to Sequence Mining
Basic Definitions of Sequences
Frequent Sequence Patterns
Desirable and Undesirable Sequences
Examples and Applications of Sequence Data

To get a good experience with sequence mining, we recommend a separate sequence mining course be added to the curriculum.

Conclusion

Sequence data mining has multitude of potential applications in diverse disciplines, from aviation safety to health care and from student management to consumer behavior management. Careful observation and deductive reasoning may provide the basis of developing an algorithm to develop software that can be used in interrupting undesirable sequence and enhance and augment desirable sequence. In this paper, a few areas of application were presented very briefly. Also, a simple example was included that gives an idea for detecting anomalous sequences.

It is not difficult to name many other examples of sequence data mining. It is important to realize that sequence mining is very practical in our lives, which makes it attractive for many researchers and developers.

References

1. Sanati, et., al, Sequence data mining for Adverse Event Detection and Action Recommendation, American Society for Engineering Education, June, 2014.

2. http://www.britannica.com/EBchecked/topic/1056150/data-mining.

3. Isinkaye O. Flasade, Computational Intelligence in Data Mining and Prospect in Telecommunication Industry, Journal of Emerging Trends in Engineering and Applied Science, (JETES) 2 (4): pp. 601-605.

4. Mabroukeh NR, Ezeife CI. A taxonomy of sequential pattern mining algorithms. ACM Computing Surveys 2010; 43:1.

5. Han J, Cheng H, Xin D, Yan X., Frequent pattern mining: current status and future directions. Data Mining and Knowledge Discovery 2007; 15 (1): 55–86.

6. Han et. al., Data Mining Concept and Technology, Third Editio, Morgan Kaufmann, 2012.

7. Robert, L. Kruse and Alexander J. Ryba, Data Structures and Program Design in C++, Prentice Hall,1999.

8. Dugdale DC, Henochowicz SI, Zieve D. Allergic reactions. ADAM Medical Encyclopedia (2013). http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001076/ (retrieved Jan 2014).

9. Nikunj, C. Oza, Data Mining for Aviation Safety, American Statistical Association, San Francisco Bay Area Chapter, October 21, 2010.

10. http://en.wikipedia.org/wiki/Weka_(machine_learning).

11. Guozhu Dong and Jian Pei, Sequence Data Mining, Springer, 2007.