Discovering Constraint-based Sequential Patterns from Medical Datasets

M. Y. Alzahrani, Fokrul A. Mazarbhuiya

Abstract: ***The problem of mining sequential patterns from medical data has received a lot of attention as it aims to discover the causal relationship between different diseases or symptoms that are present in the patient's body. Medical data contains the records pertaining to the information of the diseases or the symptoms of the patients besides the patients' personal information. The records are ordered in accordance with the time and date of the patients' visit in the hospital. Such data may offer precious information related to the cause and effect of a disease on the human body. Although, the date and time gives us chronological ordering of the occurrence of the diseases in the human body, it does not provide the information about the time intervals within which the successive diseases may occur. If the time gap of cause and effect is found to be too large, the concerned sequential pattern would be un-realistic. Considering, the time attributes of medical data, we try to address the above-mentioned problem on the sequential patterns. In this paper, we propose a method of extracting sequential patterns from medical dataset, with time-restrictions. The method extracts all sequences of diseases which occur within user-specified time intervals. The efficacy of our method is established with an experiment conducted on real life medical datasets.***

Keywords: ***Data Mining, Sequential Patterns Mining, Constraint Sequential Pattern Mining, Frequent Sequence, Maximal Sequence, Frequent Sequence within time intervals, Disease, Set of Symptoms, Frequent diseases.***

I. INTRODUCTION

Efficient mining of patterns from huge data sets is considered a complex problem in database research. Temporal data mining [1] is proposed as a notable improvement to traditional data mining. There are primarily two broad directions of temporal data mining [2]. The first direction is concerned with finding of causal relationships among temporally oriented events. The sequences are constructed using the ordered sequence of events where the cause of an event always occurs before it. The other direction is concerned with the discovery of similar patterns inside the similar time sequence or among dissimilar time sequences. The former problem is termed as sequence mining and is defined as the finding of frequent sequential patterns in the temporal datasets. *Agrawal* and *Srikant* [3] discussed the method of discovering such patterns. In [4], authors have presented, a well-described method for mining frequent sequences considering a variety of syntactic constraints like length or width limitations on the sequences, least or highest gap constraints, on the successive sequence elements.

Revised Manuscript Received on November 15, 2019

***** Correspondence Author

Mohamed Y. AlZahrani, Department of Information Technology, Department of IT, AlBaha University, KSA, Email: imohduni@gmail.com

Fokrul Alom Mazarbhuiya*, Department of Mathematics, School of Fundamental and Applied Sciences, Assam Don Bosco University, Assam, India, Email: fokrul_2005@yahoo.com

Pattern mining from medical data is one of the most vital data mining problems and recently attracted new researchers to work in this field. In [5], works have been done on extraction of temporal rules from such data. Medical data has all the historical records of information of diseases of different patients besides their personal information. Every patient has a file which contains the patient's personal information along with a set of list of symptoms or diseases from the date of their first visit in the hospital to the date of last visit. The set of list of diseases is arranged in accordance with the time or date of visits in the hospital. A sequential pattern tells us "which set of diseases follows which other set of diseases", in a specified number of patients which in turn finds out the causal relationship between different diseases. Although the above-mentioned sequential patterns give us the causal relationship between different diseases but it does not provide us the time gap of occurrence of cause and its effect. If the time interval or gap between cause and its effect is too large, the concerned sequential patter will be un-realistic and will need to be discarded.

In this paper, we address the issue in detail and propose a method of discovering sequential patterns from medical datasets where every pattern is a. time-restricted sequential pattern. Mining such sequential patterns from medical data will be helpful for a medical consultant to recognize "the causes and effects", of a set of diseases on another set of diseases. It will be also helpful for government or non-governmental agencies to formulate policies regarding the precautionary measures in a particular area before the mass outbreak of a particular disease.

The paper is prepared as follows. In section-2, we briefly talk about related work. In section-3, we introduce the Definitions and Notations used in the paper. In section-4, we discuss about the proposed algorithms. In section-5, we discuss the experimental detail of this paper and in section-6, conclusion is given.

II. RELATED WORK

The sequence mining problem, an important data mining problem has been addressed in [3, 4, 6, 7]. *Agrawal* and *Srikant* [3] have not only proposed the problem in 1995 but also established three algorithms for finding such patterns. Out these three algorithms, two unearth only maximal sequence but the third one known as A-prioriAll, unearths all patterns. In [8], an alternative algorithm has been discussed which takes into consideration the smallest and largest gaps,

as well as sliding windows. The algorithm discussed in [8] is known as GSP and it is 20 times faster than others.

Mining Medical data has been discussed widely in [5, 9]. A problem quite similar to medical data mining, the temporal frequent itemsets mining has been studied in detail in [10, 11]. The method of mining temporal rules from Medical data has been discussed in [5]. In [7], an algorithm for mining sequential patterns from medical data has been proposed which considers the time constraint. The algorithm discussed in [7] is quite similar to GSP [8]. Medical time series data mining has been studied in detail in [12]. In [13, 14], authors have suggested the techniques of association rules mining from medical image data and multiple multi-dimensional time series data respectively. In [15], an efficient constraint-based sequential pattern mining is discussed which used dataset filtering techniques. In [16], the authors discussed pattern growth method for discovering constraint-based sequential patterns. An algorithm for mining sequential patterns with time-constraint is discussed in [17] where constraint can be pushed up and down effectively.

III. DEFINITIONS AND NOTATIONS

In this section, we introduce some *terms*, *definitions* and *notations* used in the proposed algorithm.

A. Sequence

We know that any sequence is an ordered list of items (diseases). Here we redefine the sequence as the ordered list of 2-tuple consisting of diseases and time-stamps where each disease is associated with a time-stamp which is actually the time when the disease occurred. A sequence *s* is denoted as $((d_1,t_1)\rightarrow(d_2\ t_2)\rightarrow\ldots\rightarrow(d_{q,}\ t_n))$, where d_i is a disease and t_i is the time of occurrence of the disease d_i. We term a sequence as *k*-sequence, if $|\ d_i| = k$. So every sequence will be associated with a time interval which is the time between the first and last members in the sequence. For example each interval $t_{ij}=[t_i, t_j]$ contains a sequence (*d*i, *d*j) means that *d*i and *d*j were occurred within the interval t_{ij}.

A subsequence is termed as sub-collections from the sequence having same order. We can obtain a subsequence a sequence simply by removing some diseases or disease set from the sequence. A sequence *s'* contains another sequence *s* if *s* is a subsequence of s'. Similarly a sequence *s'* supports another sequence *s* if *s* is a subsequence of *s'* and the time gaps of consecutive diseases of s are within a limit.

Let *D* be the input sequence datasets in which a sequence is a set of temporally ordered transaction and each transaction is in the form $((d_1,t_1)\rightarrow(d_2\ t_2)\rightarrow\ldots\rightarrow(d_{q,}\ t_n))$.

Frequency of a 2-sequence

A counter is maintained for every sequence (a→b)*, to count support value which is zero initially. If an input sequence s in D contains the sequence* (a→b)*, then the length of the corresponding time interval say* $[t_{sa}, t_{sb}]$ *of* (a→b) *in* s, *i.e.* $/t_{sb} - t_{sa}/$ *is computed and if it is found to be less than or equal to some user-specified threshold (say* minthd*), then the support value (counter) for* (a→b) *is increased. Every input sequence in* D *supporting* (a→b) *will contribute to its support value. If the total support value for* (a→b) *is greater than or equal to some user-specified threshold (say* σ*), then* (a→b) *is said to be frequent.*

B. Frequency of an n-sequence

While finding the frequency of the *n*-sequence, $s=((a_1,t_1)\rightarrow(a_2,t_2)\rightarrow\ldots(a_{n-1},t_{n-1})\rightarrow(b_{n-1},t'_{n-1}))$, constructed from two frequent(*n*-1)-sequences $((a_1,t_1)\rightarrow(a_2,t_2)\rightarrow\ldots\rightarrow(a_{n-1},t_{n-1})$ and $((b_1,t'_1)\rightarrow(b_2,t'_2)\rightarrow\ \ldots\rightarrow(b_{n-1},t'_{n-1}))$ using candidate generation method, we proceed as follows. For every *n*-sequence, a counter is kept to count its support value which is zero initially. If an input sequence *s'* in *D* contains the sequence $s=((a_1,\ t_1)\rightarrow(a_2,\ t_2)\rightarrow\ldots\ \rightarrow(a_{n-1},\ t_{n-1})\rightarrow(b_{n-1},\ t'_{n-1}))$, then the length of the time interval (time–gap) of last 2-sequence $((a_{n-1},\ t_{n-1})\rightarrow(b_{n-1},\ t'_{n-1}))$ in *s'* is computed. If the time-gap is within some user's specified minimum threshold (*minthd*), then the support(*s*) will be increased. Here, we do not consider the time gap (intervals) of any other 2-sequence of $s=((a_1,\ t_1)\rightarrow(a_2,\ t_2)\rightarrow\ldots(a_{n-1},\ t_{n-1})\rightarrow(b_{n-1},\ t'_{n-1}))$ because all its subsequence upto (*n*-1) are already found to be frequent. As in the case of frequent 2-sequence every input sequence supporting $s=((a_1,t_1)\rightarrow(a_2,t_2)\rightarrow\ldots(a_{n-1},t_{n-1})\rightarrow(b_{n-1},t'_{n-1}))$ will contribute in the support value of *s*. If the total support is greater than or equal to some user-specified threshold (say σ), then we say $s=((a_1,t_1)\rightarrow(a_2,t_2)\rightarrow\ldots(a_{n-1},t_{n-1})\rightarrow(b_{n-1},t'_{n-1}))$ is frequent.

The sequence will give us the order of diseases according to the time of their occurrences. The frequency of 2-sequences will give us simple "cause and effects", of one disease on another. It is to be mentioned here that each 2-sequence will be associated with time interval within which the cause and effect must occur. The definition is extended for *n*-sequences. The definitions and notations discussed in this section are used in next section.

IV. METHODS PROPOSED

In this section we give the method of finding sequential patterns from medical data. In this method, we consider the time content explicitly. Here each disease / disease set will be associated with a time-stamp which is the time of occurrence of the disease. The method is described below.

The process of discovering interval-based (constraint) sequential patterns (diseases) from medical data has the following steps. The medical dataset (*D*) contains the detail information about the patients all of these are not desired in our applications. For example, only patient's dates of visits in the hospital, reports of pathological investigations (containing the presence symptoms), are important for us. Other details of patients should be removed. So, first of all we apply data pre-processing to adapt the medical data into a form appropriate for our application as done in the method. Here every transaction (record) is symptoms or diseases set along with a list of time–stamps where each disease associated with a time-stamp (time of occurrence of the disease. Thus *D* will become a dataset of disease sequences which will be input to our algorithm. Here each patient will have a set of transactions (records) and each transaction will contain one or more diseases along with time of occurrence. We call it temporal sequence dataset because of its explicit time contents.

The algorithm is multi-pass algorithm which makes multiple passes on medical data.

The first pass finds out the frequent diseases by counting support (the method of finding support is given section-3). After the pass1, we will be having frequent 1-sequence i.e. sequence of size-1. Obviously a disease is a collection of 1 or more symptoms. The algorithm for finding frequent 1-sequence is given below.

Algorithm1

Algorithm to find $F_s[1]$= the frequent 1- sequence.
C1={(d[i],t[i]); i=1, 2, ...n} where d[i] is the kth disease and t[i] the time of occurrence of d[i].
for (i=1; i≤n; i++)
set $d_{count}[i]=0$
for each input transaction) d in the dataset
for(i=1; i≤n; i++) do
{ if ($d[k]\subseteq d$) then
$d_{count}[i]$++
}
else
if $d_{count}[i]\geq\sigma$)
Add d[i] to $F_s[1]$

A count support count d_{count} is maintained for every diseases / disease set and initially $d_{\text{count}}=0$. If a disease /disease set is existed in a transaction, then d_{count} for that disease / disease set is increased by 1. For a disease / disease set if d_{count} is found to be more or equal to some pre-determined threshold value then it will be a frequent 1-sequence (frequent disease / set of diseases). The process will be repeated for each disease present in the datasets.

After that we generate candidate 2-sequence by candidate generation procedure. Then we execute the algorithm again to find the frequency of 2-sequence and so on (the definition of the frequency of 2-sequency is given in section-3). The pseudo code for the algorithm for finding *k*-sequence is as follows.

Algorithm 2

Algorithm to find sequence of size-k.
F_s1=the set of 1-sequence (disease).
k=2 do while $F_s(k-1)!=\phi$
Gen_candidate_sequences C_sk= candidate k-sequences set, where k-sequence is a sequence of diseases of the form $s=((d[1],t[1])\rightarrow(d[2],t[2])\rightarrow....\rightarrow(d[k],t[k])$ and each d[i] is a collection of one or more diseases and t[i] time of occurrence of d[i].
Prune (C_sk)
for all input sequence t in the dataset D
do
increment count of $s\in C_sk$ if ($(s\subseteq t)$& & $\lvert t[k-1]-t[k]\rvert\leq minthd$)
$F_sk=(s\in C_sk$; frequency(s) $\geq(\sigma)$)

The above algorithm gives all the time interval-based frequent sequences. Each frequent sequence will be associated with a time interval in such a way that both "cause and effect", must occur within that interval. It has two procedures i) Candidate Generation and ii) Pruning. For next level candidate generation**,** we join any two frequent sequences for which the last but previous diseases / disease set are different. For pruning, we check whether all subsequence with same starting disease (disease set) of the candidate are present in the previous level or not. We also check the time gap between last two members of the sequence is within the specified limit or not. If either of the cases is not satisfied, then it will be pruned. The candidate generation procedure and pruning procedure for the algorithm are explained below.

Candidate_sequence_generation with given Fsk-1

gen_candidate_sequences(Fk-1)
Csk:= ϕ
for all (k-1)-sequences $d_{k-1}\in F_{sk-1}$
for all (k-1)-sequences $d'_{k-1}\in F_{sk-1}$
if
$(d_{k-1}[1].t_{[k-1]}[1])=(d'_{k-1}[1],t'_{[k-1]}[1])\rightarrow(d_{k-1}[2],t_{[k-1]}[2])=(d'_{k-1}[2],t'_{[k-1]}[1])$
$\rightarrow....\rightarrow(d_{k-1}[k-2],t_{[k-1]}[k-2])=(d'_{k-1}[k-2],t'_{[k-1]}[k-2])$
$\rightarrow(d_{k-1}[k-1],t[k-1])\neq(d'_{k-1}[k-1], t_{[k-1]}[k-1])$
then
$d_k=((d_{k-1}[1],t_{[k-1]}[1])\rightarrow(d_{k-1}[2],t_{[k-1]}[2])\ldots(d_{k-1}[k-2],t_{[k-1]}[k-2])\rightarrow$
$(d_{k-1}[k-1],t_{[k-1]}[k-1])\rightarrow(d'_{k-1}[k-1], t'_{[k-1]}[k-1]))$
and $d'_k=((d_{k-1}[1],t_{[k-1]}[1])\rightarrow(d_{k-1}[2], t_{[k-1]}[2])\ldots(d_{k-1}[k-2], t_{[k-1]}[k-2])\rightarrow$
$(d'_{k-1}[k-1], t'_{[k-1]}[k-1])\rightarrow(d_{k-1}[k-1],t_{[k-1]}[k-1]))$
$C_{sk}:=C_{sk}\cup\{ d_k, d'_k\}$

Pruning

prune(C_{sk})

for all $d_k\in C_{sk}$
for all (k-1)-subsequences $d_{k-1}$of d_k having same starting disease
do
if $d_{k-1}\notin F_{k-1}$
then $C_k:=C_k\setminus\{d_k\}$

The algorithm supplies all frequent sequence where each frequent sequence is an ordered list of diseases / diseases associated with list of time interval which is the interval of occurrence of consecutive diseases (disease set) in the sequence. For instance, if $(D_c\rightarrow D_e)_{[t,\ t']}$ is frequent, then we say, "D_c and D_e occurred frequently in the time frame [*t*, *t*/] and D_e follows D_c", Then D_c can be considered as cause and thus D_e can be its effect within the time interval [*t*, *t*/]. The proposed system architecture is given fig-1below.

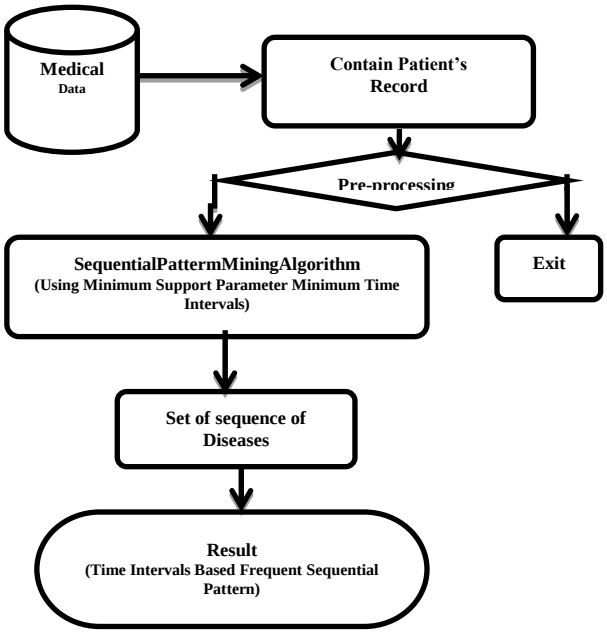

Fig.1: System architecture

The proposed architecture is explained as follows. The Medical Database contains all the patient's records of information regarding diseases along patient's personal details like name, address, Ph. No., doctor's name and date / time of patient's visit in the hospital etc. Each record of disease contains the detail about of presence or absence symptoms of the diseases along with date/ time of occurrence of disease / diseases and it is associated to a single visit of a patient. It is to be mentioned here date / time of occurrence of disease is same as patient's visiting date / time. So a patient may have multiple records ordered according to the time of

his or her visit in the hospital. All other information is undesirable. First of all our system takes Medical Data as input and removes all undesirable information convert it to the dataset of patient's records. Patient's each records contains symptoms of the diseases. The records are pre-processed one by one to convert it to a set of diseases by eliminating the missing values and equating each set of symptoms into disease / diseases. Then we will have set of sequence of diseases where each sequence is a actually a sequence of 2-tuple containing diseases and their time of occurrences and is associated to a single patient. So we may have as many sequences as the number of patients. This set is input to the SequetialPattenrMiningAlgorithm. The algorithm has two user's specified parameters namely minimum support parameter and minimum time interval. The minimum support parameter will filter the sequences based on its frequency of occurrence and minimum time interval will filter those sequences whose length of occurrence are more than the user-specified length. Finally our system gives all the frequent sequences which occur within user's specified time-interval.

V. EXPERIMENT AND ANALYSIS OF RESULTS

For the experiment, two different datasets have been used which are collected from UCI machine learning repository. A brief description of the two datasets is given in Table-I below.

TableI: Description of Datsets

Dataset	Dataset charc	Attribute charc	No. of Insta	No. of Attr	Miss. Val	Area	Donatn Date
BUPA	Multivariate	Categorical, Integer, Real	345	7	No	Life	15/05/1990
Heart Disease	Multivariate	Categorical, Integer, Real	303	15	Yes	Life	01/07/1988

Bupa dataset [18]: This dataset contains 345 single male patients having 6 numeric attributes. Out of these six, five attributes corresponds to blood tests which are thought to be relevant to liver disorders. The last attribute corresponds to the number of alcoholic beverages drunk per day. The dataset is having two classes.

Heart Disease dataset [19, 20]: This dataset has 13 attributes and 2 classes. It has a total 270 records.

We take $\sigma = 0.4$ and *minthd*=15 days. We also determine the dataset sizes by the number of patient's electronic records and apply the algorithm for such datasets. A partial view of the detail results is explained using table-II, fig.2, fig.3 and fig.4.

The table-II gives the comparative study of our method for two different datasets mentioned above which is presented graphically in fig-2 and then using bar diagram in fig-3. The blue line indicates the results from Bupa dataset and red line indicates that from Heart Disease dataset. The datasets sizes (number of patient's records) are represented by x-coordinates and number of frequent sequence are represented by y-coordinates. From the above figures, we can conclude that the method has extracted quite similar kind patterns from both the datasets. The fig-4, is graphical representation of number of frequent sequences obtained from above-mentioned datasets with respect to the size of time intervals (user's specified) of occurrence of diseases in the sequence. The figure indicates that the number of frequent sequences increases with the increase in the sizes of time intervals. However the rate of increase in case of Heart Disease dataset is little bit higher than that for Bupa dataset.

Table-II: Number of frequent sequence for different sizes of datasets

Dataset sizes (Number of patient's record) (BUPA, Heart Disease)	Number of sequential patterns by our method	
	BUPA	Heart Disease
0	0	0
5	1	2
10	3	4
50	4	5
100	4	6
200	5	6
Full dataset	6	7

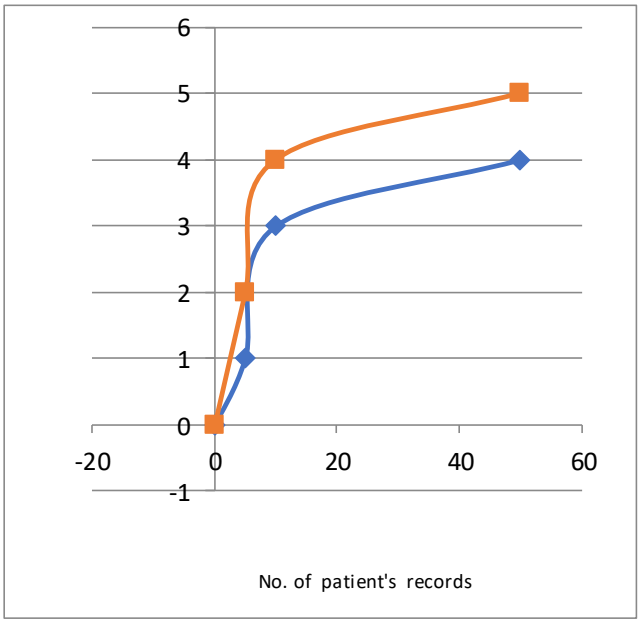

Fig-2: Frequent sequence vs sizes of datasets

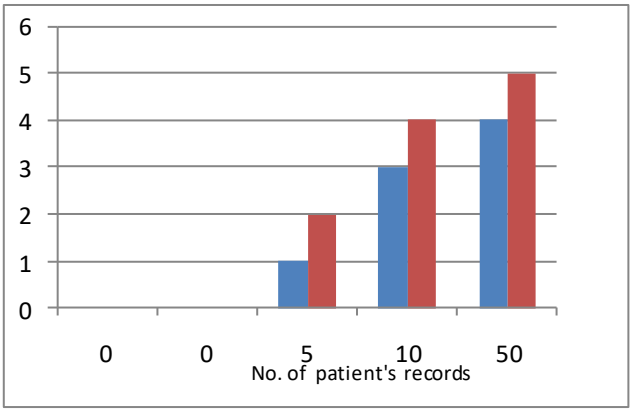

Fig-3: Frequent sequence vs sizes of datasets

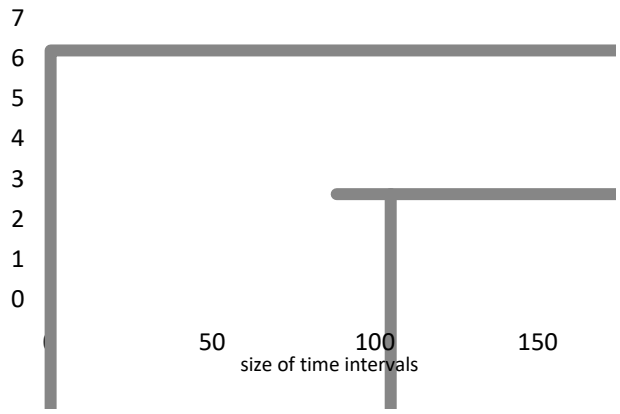

Fig-4: Frequent itemser vs sizes of time intervals

VI. CONCLUSION

In this paper, we propose an algorithm for finding sequential patterns from medical datasets. The algorithm is used for finding time-constraint sequential patterns. Our algorithm extracts all those sequential patterns for which the time gaps of "causes and their effects", are within specified limit. The algorithm follows level-wise approach. It starts with extracting frequent 1-sequences and goes on extending the sizes of the sequences till either no extension is possible or a particular level is empty. As we mentioned earlier that Medical data contains all the health related information about any patient right from the starting of his visits in the hospital and the records are arranged and kept in the order of the time of visit of the patient. The obtained sequential patterns will be useful in determining disease or diseases that follow other disease or diseases in a specified number of patients within a specified time frame. For example, if a disease or disease set D_2 follows another disease or diseases set D_1 in a specified number of patients and they occurred within a specified time frame, then the cause of D_2 would be a D_1 and effect of D_1 would be D_2. The algorithm discussed here is basically a level-wise algorithm. Finally, we have implemented the algorithms using two standard datasets and have tested the algorithm's efficacy.

REFERENCES

1. C. M. Antunes, and A. L. Oliviera, "Temporal Data Mining an overview", Workshop on Temporal Data Mining-7th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, (2001).
2. J. F. Roddick, and M. Spillopoulou, "ABiblography of Temporal, Spatial and Spatio-Temporal Data Mining Research", ACM SIGKDD, (1999).
3. R. Agrawal, and R. Srikant, "Mining sequential patterns", In Proc. of 11th Int'l Conf. on Data Engineering, IEEE, pp.3-14, (1995).
4. M. J. Zaki, "Efficient enumeration of frequent sequences", In 7th Int'l Conf. on Information and Knowledge Management, (1998).
5. H. Meamarzadeh, M. R. Khayyambashi and M. H,Saraee, "Extracting Temporal Rules from Medical data". IEEE, pp. 327-331, (2009).
6. H. Manilla, H. Toivonen and I. Verkamo, "Discovery of frequent episodes in event sequences", Data Mining and Knowledge Discovery: An International Journal 1(3), (1997), pp. 259-289.
7. M. J. Zaki; "Sequence Mining in Categorical Domains Incorporating Constraints", CIKM 2000, Mc. Lean, VA, USA, pp. 422-429, (2000).
8. R. Srikant and R. Agrawal; "Mining Sequential Patterns: Generalizations and Performance Improvements", In Proc. of the 5th International Conference on Extending Database Technology (EDBT'96), Springer-Verlag, London UK, (1996).
9. F. A. Mazarbhuiya, and M. Y. AlZahrani; "An efficient method for clustering periodic patterns", Computing Conference 2017, London, U. K, (2017)
10. A. K. Mahanta, F. A.Mazarbhuiya and H. K. Baruah, "Finding Locally and Periodically Frequent Sets and Periodic Association Rules", PReMI'05, LNCS 3776, 576-582, (2005).
11. A. K. Mahanta, F. A. Mazarbhuiya, And H. K. Baruah, "Finding Calendar-based Periodic Patterns", Pattern Recognition Letters, Vol. 29(9), Elsevier publication, pp. 1274-1284, (2008).
12. C. Catley, H. Stratti and C. McGregor, Multi-Dimensional Temporal Abstraction and Data Mining of Medical Time Series Data: Trends and Challenges, In proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2008, pp. 4322-4325, (2008).
13. A. Olukunle and S. Ehikioya, "A Fast Algorithm for Mining Association Rules in Medical Image Data", IEEE. p1-7, (2002).
14. G. N. Pradhan and B. Prabhakaran, "Association Rule Mining In Multiple, Multidimensional Time Series Medical Data", IEEE. pp.1-4, (2009).
15. Tadeusz Morzy, Marek Wojciechowski, and MaciejZakrzewicz, "Efficient constraint-based Sequential pattern mining using Dataset filtering techniques", Databases and Information Systems II**,** Fifth International Baltic Conference, Baltic DB&IS'2002 Tallinn, Estonia, pp. 297-309, (2002).
16. Jian Pei, Jiawei Han, and Wei Wang, "Constraint-based sequential pattern mining: the pattern-growth methods", Journal of Intelligent Information Systems, Vol. 18, Issue 2, pp. 133-160, (2003).
17. Anita Zala and Mehul Barot, "Mining Sequential Pattern with Time-Constraint", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 7, (2013).
18. McDermott & Forsyth, "Diagnosing a disorder in a classification benchmark", Pattern Recognition Letters, Volume 73, (2016).
19. R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, and V. Froelicher; "International application of a new probability algorithm for the diagnosis of coronary artery disease", American Journal of Cardiology, 64, (1989), 304--310.
20. J. H. Gennari, P. Langley, and D. Fisher; "Models of incremental concept formation". Artificial Intelligence, 40, (1989), 11--61.

AUTHORS PROFILE

Dr.Mohammed Y. Alzahrani has received his B.Sc. in Computer Engineering from Albaha Private College of Science and M.Sc. in Information Technology from Heriot Watt University, Edinburgh, UK. After this he has obtained Ph. D. degree in Computer Science from Heriot Watt University, Edinburgh, UK. Currently he is working as the Dean of College of Computer Science and Information Technology at Al Baha University, Al Baha, Kingdom of Saudi Arabia. His research interest includes Model Verification, Data Mining and Information Security.

Dr. Fokrul Alom Mazarbhuiya received his Ph.D. degree in Computer Science from Gauhati University (2007), India. He had been working in the College of Computer Science and IT at King Khalid University, and then Albaha University, Kingdom of Saudi Arabia since 2008 to 2018. He is currently working in Department of Mathematics, School of Fundamental and Applied Science, Assam Don Bosco University, India. His research interest includes Data Science, Information security, and Fuzzy logic.