ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

A Survey on Medical Data Mining for Periodically Frequent Diseases

Mohammed Abdul Khaleel[1], Sateesh Kumar Pradhan[2], G.N.Dash[3]

Research Scholar, Sambalpur University, India[1]

Post Graduate department of Computer Science, Utkal University, India[2]

Post Graduate, Department of Physics, Sambalpur University, India[3]

Abstract: Medical data mining is the field of research pertaining to health care domain which helps in capturing latent relationships among attributes. This kind of knowledge discovery in modern world has become indispensable as clinical data is very huge and manual interpretation is not possible. Mining transactional databases containing medical data has plethora of real world utilities such as medical diagnosis, expert decision making and so on. Medical data mining also helps in promoting evidence based research that is essential and suitable for health care industry. Though there are many areas of research in medical data mining, in this paper, we focus on the data mining approaches or methods that help in finding periodically frequent diseases. A pattern which occurs at regular time intervals is said to be periodically-frequent. Periodic frequent pattern discovery can help people concerned to make well informed decisions. Towards this end, we analyze various methods that are used in medical data mining for discovering actionable knowledge pertaining to periodically frequent diseases.

Index Terms: Data mining, medical data mining, frequent patterns, periodically frequent diseases

I. INTRODUCTION

Periodic frequent patterns are the trends in the data that occur frequently in regular intervals. A pattern is nothing but set of items that exhibit an interesting fact. Often that interesting fact helps enterprises to make expert decisions. Usually periodically frequent patterns are extracted with constraints such as minimum support and maximum periodicity. The minimum support helps to filter the patterns that do not occur with given frequency while the periodicity criterion can help to control the maximum time difference between two occurrences of pattern in the given data set. This kind of medical data mining can bring about rate items that are of interest to discover knowledge [1].
In this paper our contributions are described here.

- Survey of medical data mining techniques that are used for mining periodically frequent diseases in health care domain.
- Analyzing the present state-of-the-art pertaining to medical data mining with respect to extracting periodically frequent patterns.

The remainder of the paper is structured as follows. Section II provides review of literature pertaining to medical data mining, especially on mining periodically frequent diseases. Section III analyzes the present state-of-the-art on mining periodically frequent patterns. In other words this section provides summary of findings with discussion in some detail. Section IV provides conclusions and directions for future work.

II. RELATED WORKS

This section reviews literature on mining periodic frequent patterns. It throws light into mining techniques used on transactional databases for discovering periodically frequent patterns. Surana et al. [2] proposed a data mining method named "MaxCPF-Tree" for mining periodic frequent patterns. The algorithm is made with multiple constraints pertaining to periodicity and minimum support. Two important parameters considered in the algorithm are "minsup" and "maxprd". The values of these parameters determine the output of the algorithm. Sample transactional database is used for experiments. However, this algorithm can be explored for medical data mining as well.

TID	Items
1	a, b, h
2	c, d
3	a, b, d
4	c, e, f
5	a, b

TID	Items
6	c, d, g
7	a, b, c, e
8	d, e, f
9	a, b, c
10	g, h

Table 1

Pattern	S	P	I	II	III
{a}	5	2	✓	✓	✓
{b}	5	2	✓	✓	✓
{c}	5	2	✓	✓	✓
{d}	4	3	✓	✓	✓
{e}	3	4	✓	✓	✓

Table 2

Pattern	S	P	I	II	III
{f}	2	4	✓	✓	✓
{a, b}	5	2	✓	✓	✓
{c, d}	2	4	✓	✓	X
{c, e}	2	4	✓	X	X
{e, f}	2	4	✓	✓	✓

Table 3

Figure 1 – Tables containing transactional data and mined patterns [2]

Table 1 shows the transactional database which contains unique transaction IDs and corresponding items. Table 2 shows various patterns with values given to the parameters "minsup" and "maxprd" denoted by S and P respectively. Columns I, II, and III represent the periodically frequent patterns that have been mined. From the experiments the researchers came to know that in order to obtain rare items and also frequent patterns it is essential to provide low minimum support value and high periodicity value.

Chen and Liu [3] proposed a mining algorithm for finding frequent patterns in biological sequence. They introduced a concept known as primary pattern and then constructed

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

prefix tree for performing mining for frequent primary patterns. They algorithm is named "FBPM" which improve mining performance and provides accurate results. The notion of primary pattern is illustrated using the data in Table 4.

S_x	loc
a	4
ab	9
abcb	5
ac	2
b	10
ba	8
bacaa	1
bc	6
caab	3
cbab	7

Table 4 – Illustrates primary patterns [3]

As can be seen in Table 4, the results reveal the primary patterns for S. These primary patterns are sorted before using them for further mining process and a prefix tree is generated for the primary patterns of S. Figure 1 shows the generated prefix tree representing primary patterns of S.

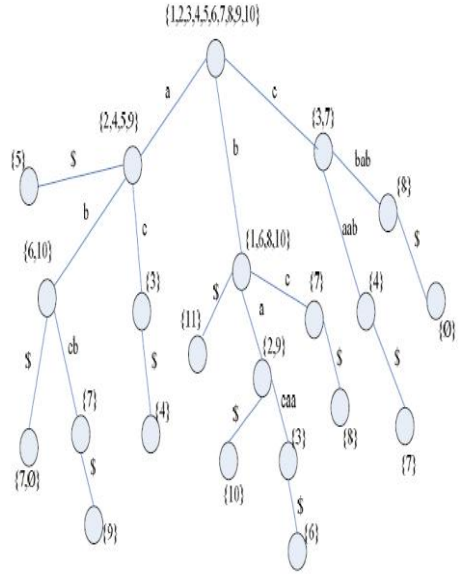

Figure 1 – Prefix tree for representing primary patterns of S [3]

After construction of prefix tree, mining is performed on the tree structure that extracts periodically frequent patterns. The prefix tree also helps in avoiding irrelevant patterns. The algorithm starts from root node and moves to all nodes layer by layer fashion. The path in the prefix tree ends at a. The similar kind of work was done by Xiong et al. [4] whose algorithm was named "BioPM" meant for protein sequence mining where multiple support values are used for mining patterns. The results of FBPM [3] are compared with that of BioPM [4]. The comparison is made in terms of computation time and vs. the number of sequence and the minimal support threshold. The results of [3] are also compared with Apriori algorithm [5]. The results are as shown in Figure 2.

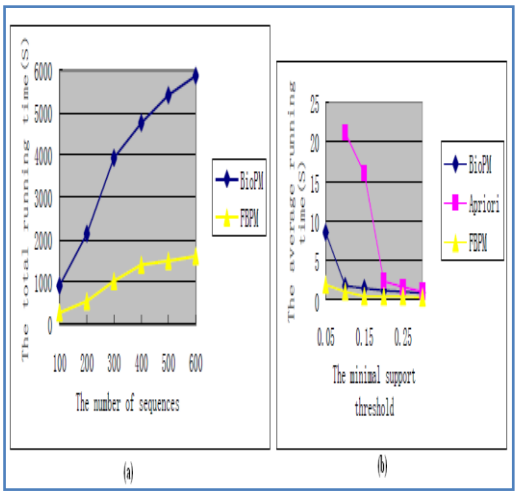

Figure 2 – Performance of FBPM compared with Apriori and BioPM [3]

As can be seen in Figure 2 (a) it is evident that the FBPM algorithm outperforms the BioPM algorithm with its performance. The total running time for various number of sequences is low for FBPM. The reason behind the performance bottleneck with BioPM is that it generates lots of intermediary and short patterns that consume more time. The performance of FBMP is more when compared with other algorithms such as BioPM and Apriori with respect to average time taken with given minimal support threshold. The results reveal that the support is minimum the average time taken is more [3]. Apriori is one of the widely used algorithms for obtaining association rules. Ilayaraja and Meyyappan [6] explored it well for medical data mining. They extracted frequently occurring diseases using Aprioi algorithm which was originally proposed by Agarawal et al. [5]. In fact the authors of [6] used temporal data mining approach that could extract month wise frequency of various diseases that affect patients. However, the algorithm proposed by them can be altered in order to obtain periodically frequent diseases.

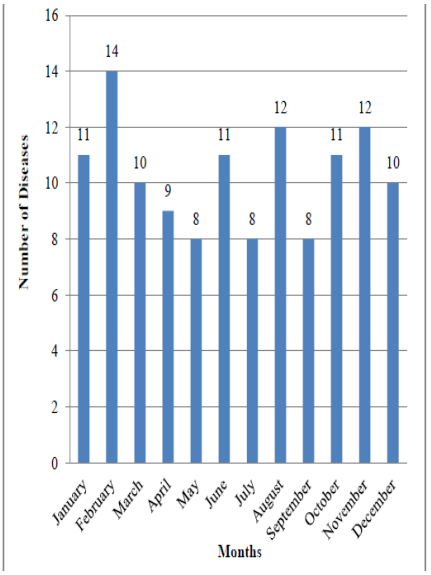

Figure 3 – Month wise occurrence of diseases [6]

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

As can be seen in Figure 3, it is evident that the algorithm has mined the medical data and visualized the periodically frequent patterns. In other words the results reveal the number of diseases that occur in each month of the given year.

Sridevi and Rajaram [7] proposed a new method for extracting periodic frequent patterns from transactional database. Their approach makes use of time stamp in order to find the time interval and make decisions while presenting data. The processing is done in three phases. In the first phase dataset is taken and the data is identified for processing. In phase II transitional pattern mining is done. In phase III Allen's algebra is used to find periodic frequent patterns. Afterwards their method is compared with other existing methods. Their framework is as shown in Figure 4.

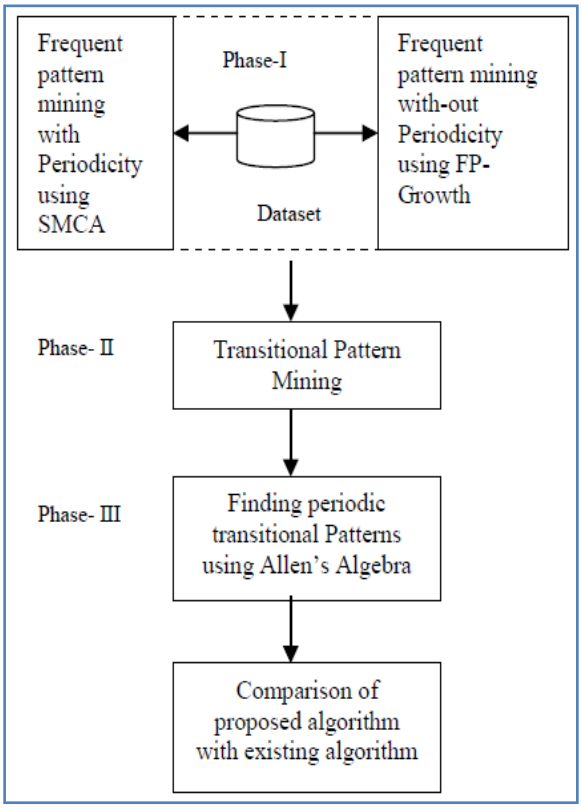

Figure 4 – Overview of the Framework for Efficient Periodic Transitional Patterns [7]

As can be seen in Figure 4, the framework makes use of algorithms in order to mine frequent patterns. The algorithms used are FP – Growth [8] and SMCA [9]. The former is used to mine frequent patterns with periodicity while the latter can be used without periodicity. Other mining algorithms for mining periodic frequent patterns on time series database include Singular Periodic Pattern Mining (SPMiner) [9], Asynchronous Sequence Mining (APMiner) [9], Complex Periodic Pattern Mining (CPMiner) [9] and Multievent Periodic Pattern Mining (MPMiner) [9].

Huiping Cao et al. [10] defined the problem pertaining to periodic patterns in a database containing spatiotemporal data. For retrieving maximal periodic patterns they proposed an effective algorithm. Catley et al. [11] explored mining patterns on medical data that are temporally abstracted. They applied their mining technique to multi-dimensional clinical data. They could extract trends from clinical data and classify them as data, results, knowledge, and integration. The overview of the trends extracted is presented in Figure 5.

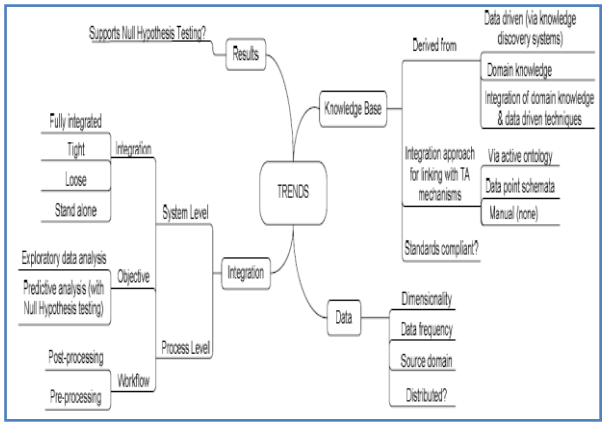

Figure 5 – Overview of the patterns obtained from clinical data [11]

As can be seen in Figure 5, the trends classified include multi-dimensional, high frequency data, real world clinical data, supporting null hypothesis testing, process level integration, synthesized knowledge base, and system level integration. Berlingerio et al. [12] studied a real world medical case study in order to explore trends in temporal dimension. They applied time annotated sequences [13], [14] for extracting trends. Catley et al. [15] proposed an integrated temporal data mining approach for mining patterns on medical data. Meamarzadeh et al. [16] built an application of temporal data mining. They extracted temporal rules from medical data that can be used for making well informed decisions. They built a methodology which is based on Allen's temporal relationship theory. The temporal interval relation among the data items is used to build directed acyclic graph gestational diabetes as shown in Figure 6.

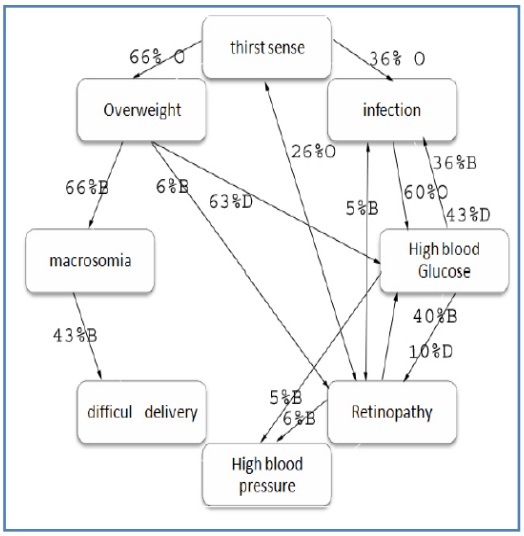

Figure 6 – The directed acyclic graph gestational diabetes [16]

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

Tsumoto et al. [17] proposed temporal data mining method for temporal knowledge pertaining to nursing practice. This research could improve services in health care domain. The temporal patterns obtained helped to make decisions in a hospital. The mining process used temporal data analysis that generates a decision tree for making expert decisions.

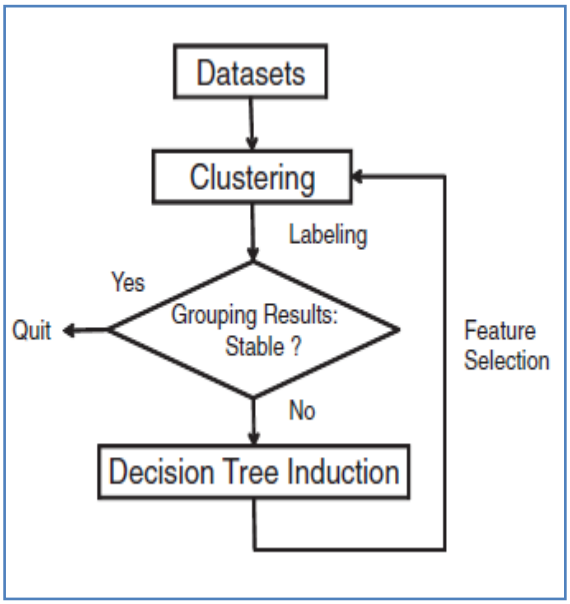

Figure 7 – Mining process [17]

The mining process is applied to medical data base where the methods used are multidimensional scaling and clustering. The dataset contains clinical data pertaining to lung cancer and cataracta. This method proved to be innovative for improving hospital services and management. Froelich and Wakulicz-Deja [18] proposed adaptive fuzzy cognitive maps for periodic frequent patterns. They extracted medical concepts from data containing temporal relationships. The trends obtained temporally include condition of patients, drugs prescribed, effects of drugs, and any side effects and so on. Thus their method was effective in knowledge representation and mining trends from that.

Spenceley and Warren [19] proposed an intelligent online interface to deal with electronic medical records. Temporal data mining approach is used by the application that which proved to be useful in predicting data requirements and has potential for further application of temporal data mining to obtain periodic frequent patterns. Lai et al. [20] presented more flexible model to extract periodic frequent patterns. The data used for experiments is time series data. Their method explored mining frequent patterns based on the temporal abstractions. The mining process employed by them also used the parameter minimum support in order to have control over the patterns. Thus the resultant patterns will be used by domain experts for decision making.

III. SUMMARY OF FINDINGS

This section provides the summary of the findings conceived from review of literature pertaining to some sort of temporal mining which aimed at producing periodically frequent patterns from medical data sets.

Surana et al. [2] proposed a data mining method named "MaxCPF-Tree" for mining periodic frequent patterns. Multiple constraints for the parameters "minsup" and "maxprd" are the important features in their method to deal with rate items problem. Chen and Liu [3] proposed a mining algorithm for finding frequent patterns in biological sequence. They used primary pattern concept for the first time. They extracted primary patterns first and then the generated data structure is sued for mining periodically frequent patterns. Ilayaraja and Meyyappan [6] explored Apriori for medical data mining. They could extract month-wise occurring diseases as part of their periodic frequent pattern mining. Sridevi and Rajaram [7] proposed a new method for extracting periodic frequent patterns from transactional database. Allen's algebra played an important role in this approach. They made use of algorithms such as FP-Growth and SMCA for achieving the results. They also used other mining approaches such as Singular Periodic Pattern Mining (SPMiner) [8], Asynchronous Sequence Mining (APMiner) [9], Complex Periodic Pattern Mining (CPMiner) [8] and Multievent Periodic Pattern Mining (MPMiner) [8] for time series datasets. Huiping Cao et al. [10] made mining on spatiotemporal data while Catley et al. [11] explored mining on multi-dimensional clinical data. Time annotated sequences technique is used in [13] and [14]. Allen's temporal theory is also used in [16]. In [17] periodic pattern mining is done to improve services in hospital with respect to nursing which deals with diseases like cancer and cataracta. Adaptive fuzzy cognitive maps are used for periodic frequent pattern mining in [18] for effectively extracting hidden periodic frequencies in clinical database. Intelligent online interface for mining electronic patient records was explored in [19] while a more flexible model for periodic frequent pattern mining is explored in [20]. From the best knowledge obtained from these insights, we are proposing a new method for effective periodic pattern mining of medical data in our future work.

IV. CONCLUSIONS AND FUTURE WORK

In this paper we studied the temporal relationship which is hidden in medical databases in terms of mining periodic frequent patterns. There are two parameters that are to be supplied by domain experts in order to extract very useful periodic frequent patterns that can be used for expert decision making. These parameters are minimum support and periodicity. However, some methods found in the literature make use of both while other methods use only support parameter. The rate item problem is also identified and which can be resolved using low support value and high periodicity as it will not ignore certain values for consideration. From this research it is understood that further work is required in order to have more effective method that can help in extracting periodically frequent patterns that can be used by hospitals in the health care industry. Towards this end, in our future work we are going to develop a tool along with a new method which will extract periodically frequent diseases which are of interest to decision makers in health care domain.

REFERENCES

[1] S. K. Tanbeer, C. F. Ahmed, B.-S. Jeong, and Y.-K. Lee. Discovering periodic-frequent patterns in transactional databases. In PAKDD '09: Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, pages 242–253, Berlin, Heidelberg, 2009. Springer-Verlag.

[2] Akshat Surana, R. Uday Kiran and P. Krishna Reddy. (n.d). An Efficient Approach to Mine Periodic-Frequent Patterns in Transactional Databases. IEEE. p1-12.

[3] Ling Chen and Wei Liu. (2011). An Algorithm For Mining Frequent Patterns in Biological Sequence. IEEE. p63-68.

[4] Xiong Yun, Zhu Yangyong. BioPM: An Efficient Algorithm for Protein Motif Mining[C]. In: Proc. of ICBBE'07. [S. l.]: IEEE Press,2007. 394-397.

[5] Srikant R, Agrawal R. Mining sequential patterns: Generalization and performance improvements[C]. In: Apers PMG, Bouzeghoub M, Gardarin G, eds. Advances in Database Technology, Proc. of the 15th Int'l Conf. on Extending Database Technology. London: Springer- Verlag, 1996: 3-17.

[6] M. Ilayaraja and T. Meyyappan. (2013). Mining Medical Data to Identify Frequent Diseases using Apriori Algorithm. IEEE. p194-199.

[7] S.SRIDEVI and Dr.S.RAJARAM. (2013). AN EFFICIENT METHOD FOR MINING PERIODIC TRANSITIONAL PATTERNS IN TRANSACTION DATABASE. JATIT. 5 (1), p74-83.

[8] Florian Verhein. (2008). Frequent Pattern Growth (FP-Growth) Algorithm. The University of Sydne. p1-10.

[9] Kuo-Yu Huang and Chia-Hui Chang, "SMCA: A General Model for Mining Asynchronous Periodic Patterns in Temporal Databases", IEEE Trans on Knowledge and Data Engg.17,2005,pp. 774-785.

[10] Huiping Cao, Nikos, and David W Cheung, "Discovery of Periodic Patterns in Spatiotemporal Sequences", IEEE Trans on Knowledge and Data Engg.Vol 19,2007,pp.453-467.

[11] Christina CatleyHeidi Stratti and Carolyn McGregor. (2008). Multi-Dimensional Temporal Abstraction and Data Mining of Medical Time Series Data: Trends and Challenges. IEEE. p4322-4325.

[12] Michele Berlingerio, Francesco Bonchi Fosca Giannotti and Franco Turini. (2007). Mining Clinical Data with a Temporal Dimension: a Case Study.IEEE. p429-436.

[13] F.Giannotti,M. Nanni, and D.Pedreschi. Efficient mining of temporally annotated sequences. In Proceedings of the Sixth SIAM International Conference on Data Mining , 2006.

[14] F. Giannotti, M. Nanni, D. Pedreschi, and F. Pinelli. Mining sequences with temporal annotations. In Proceedings of the 2006 ACM Symposium on Applied Computing (SAC), pages 593–597, 2006.

[15] Christina Catley, Kathy Smith, Carolyn McGregor and Mark Tracy. (2009). Extending CRISP-DM to Incorporate Temporal Data Mining of Multidimensional Medical Data Streams: A Neonatal Intensive Care Unit Case Study. IEEE. p1-5.

[16] Hoda Meamarzadeh, Mohammad Reza Khayyambashi and Mohammad Hussein Saraee. (2009). Extracting Temporal Rules from Medical data.IEEE. p327-331.

[17] Shusaku Tsumoto, Shoji Hirano and Haruko Iwata. (2012). Temporal Data Mining of Order Entry Histories for Characterization of Medical Practice. IEEE. p1-4.

[18] Wojciech Froelich and Alicja Wakulicz-Deja. (2009). Mining Temporal Medical Data Using Adaptive Fuzzy Cognitive Maps. IEEE. p16-23.

[19] Susan E. Spenceley and James R. Warren. (1998). The Intelligent Interface for On-Line Electronic Medical Records using Temporal Data Mining. IEEE. p1-9.

[20] Chih Lai, Nga T. Nguyen, Dwight E. Nelson. (n.d). Mining Periodic Patterns from Floating and Ambiguous Time Segments. IEEE. p1-12.