On Association Rule Mining from Diabetes Medical History

Purnomo Husnul Khotimah[1] Akihiro HAMASAKI[2] Masatoshi YOSHIKAWA[1] Osamu SUGIYAMA[3]
Kazuya OKAMOTO[4] and Tomohiro KURODA[4]

[1]Graduate School of Informatics, Kyoto University, Japan
[2]Center for Diabetes & Endocrinology, Taduke Kofukai Medical Research Institute Kitano Hospital, Japan
[3]Preemptive Medicine & Lifestyle-Related Disease Research Center, Japan
[4]Div. of Medical IT & Admin. Plan., Kyoto University Hospital, Japan

E-mail: hkhotimah@db.soc.i.kyoto-u.ac.jp, yoshikawa@i.kyoto-u.ac.jp,
{hamasaki,sugiyama,kazuya,tomo}@kuhp.kyoto-u.ac.jp

Abstract In chronic medication treatment, it is important to find pattern of adjacent medication episodes to learn the physician strategy in a longitudinal medical history. Frequent pattern mining is commonly used to extract patterns from large datasets. Traditional frequent pattern mining considers to find subsequences that are frequent. The subsequences may be consisted with non-consecutive subsets of itemset. However in chronic medication treatment the medication combination and the sequence of the treatment are considered essential. Hence, we proposed a new method for mining the full and adjacent itemsets. In current study, we visualize association rules from 1-sequence patterns generated by our mining method from diabetes medical history, to construct a medication trajectory graph. We conduct an experiment by visualizing comparing medication trajectory graph for medical history from 2000 until 2009. The results show that our proposed method is able to help the domain expert to quickly identify interesting patterns.

Keyword frequent sequence pattern, singleton mining, chronic medication treatment

1. Introduction

In chronic medication treatments, it is important to do pattern mining from adjacent treatment. As chronic diseases are illness that span the patient's lifetime and may progress over the years, the physicians need to able to manage the patients' condition so that it achieves the ideal condition and slows the rate of the illness progression. By mining adjacent treatment, then we will be able to observe the medication transition events to learn about physician's strategy in managing the patient condition over longitudinal datasets.

Diabetes is a common chronic disease with the highest prevalence number based on WHO documentation. In the case of diabetes, the selection of pharmacotherapy is considered essential [1]. The appropriate combination of medications should be selected in accordance with the patient conditions. Table I, shows the list of medication types used for diabetes. In addition, the treatment in diabetes is recommended to be stepwise, which means that the selection of the first treatment affects the next selection of treatment when the disease progresses. For example, we have the following sequence of medication transition events: <*Sulfonylurea, α-Glucosinadase Inhibitor*> → <*α-Glucosinadase Inhibitor, DPP4 Inhibitor*> → <*DPP4 Inhibitor*>. Using the traditional Apriori method, the medication pattern of <*α-Glucosinadase Inhibitor*> → <*DPP4 Inhibitor*> is considered an instance of frequent pattern because it is contained in the medication transition events. However, this pattern may have less meaning for the physicians because it is not represent the actual medications and sequence order.

Table I
Medication Types

No	Medication Type
1	Sulfonylurea (SU)
2	Rapid-acting insulin secretagogues (RaIS)
3	α-Glucosidase inhibitors
4	Biguanides
5	Thiazolidinediones
6	DPP-4 inhibitors

The traditional Apriori method is firstly developed to solve the market basket, that is to study customer behavior [2]. The algorithm is trying to find frequent items bought together by the customer. In market basket case, the item number may be very large. Hence, the possible patterns generated by such itemset may be enormous and it will consume a high computation cost. However, in the case of chronic medication treatments, the items is considerably small. In the diabetes case, there are less than 15 medication types based on [1]. Thus, it is possible to enumerates all possible combination of items. In addition, the number of possible combinations will also be fewer because there are evidence based recommendation of chronic medication treatment that physicians

used as guideline in managing the patients' condition. Other condition is that in the market basket case, items purchased by customer are considered independent. By contrast, in chronic medication treatment the medication combination may have confluence. Hence, a distinct medication combination might have different function compared to another combination. Furthermore, in the market basket case, the subsequence may be non-consecutive. This condition differs to chronic medication treatment that the selection of the first treatment affect to the selection of the next treatment. Therefore, we propose a new method for finding frequent patterns that its itemsets are the full itemsets (i.e., singleton, which is itemset that represents the actual events) and adjacent to each other (consecutive).

The proposed method is proven to be generating a more compact resultset compared to the one produced by the traditional Apriori method and the result set has different order of ranking based on its pattern support. These results has been presented in our previous paper that investigated a study case of comparing frequent pattern (FP) result set of our proposed method mining and Apriori-based mining using similarity scoring on their result set rank based on the support value [3]. In current study, we are interested to analyze deeper into the association rules result set of both method in order to give a better understanding of the proposed method's strong points. We consider this study is important because extracting interesting association rules is the final goal of frequent pattern mining activity.

In order to do so, we develop a medication trajectory from frequent patterns to visualize the association rules to help the domain expert to explore the association rules. We focus our study in the usage of weighted directed graph to develop the medication trajectory. Compared to common tree graph that depicts classification [4][5] or used for clustering [6], our graph shows a trajectory, which is aggregated from medication transition patterns.

The rest of the paper is structured in the following manner: In Section 2, we introduce the related work including physician strategy in managing chronic diseases, frequent pattern mining, singleton mining, and data visualization. Section 3 provides a method to develop a medication trajectory graph. Then, we will discuss the result in Section 4, and finally, we conclude the results in section 5.

2. Related Work

2.1. Physician Strategy in Managing Chronic Diseases

Chronic diseases are long term diseases that may lead to health deterioration and can compromise quality of life through physical limitation and disability [7]. There are no cure for chronic diseases. Therefore, the patient's condition must be managed through a comprehensive longitudinal medical care to maintain the quality of life and inhibit the disease progression.

Currently, health care provider is recommended to give evidence based treatment. Evidence based medical guideline is developed and updated by conducting clinical studies onto the physician strategy. The physician strategy in managing the patient condition is able to be observed via the patients' medical history. For example, in diabetes, a physician may start from monotherapy for sometime and change to a medication combination after the diseases progress.

Our study consider medication transition events are essential because they are kind of marking when the patients' condition changes and the physician needs to modify the treatment. We have listed medication transition events as the following five actions:

- *Add* is when new medication(s) are added to the previous medication.

Table II
Medication Type Transitions of Six Patients

pid	Medication Type Transition
1	<1,3> → <3,5> → <3,6>
2	<3,4> → <3,5> → <1,3,5> → <1,4> → <1,4,6>
3	<1> → <1,4> → <1,4,6>
4	<1>
5	<5> → <4,5> → < 4 >
6	<1> → <1,4> → <1,3,4> → <1,3,4,6>

- *Stop* is when previous medication(s) are stopped from the previous medication.
- *Switch* is when new medication(s) are added and previous medication(s) are stopped.
- *Increasing* is when the dosage of medication(s) is increased.
- *Decreasing* is when the dosage of medication(s) is decreased.

We also consider physicians' action to prescribe the same medication with the previous medication as ``continue".

Example 1: From Table I, for patient id 2, we have medication transition event ``add" in the sequence $<3,5>\rightarrow<1,3,5>$ and $<1,4>\rightarrow<1,4,6>$ and ``switch" in the sequence $<3,4>\rightarrow<3,5>$ and $<1,3,5>\rightarrow<1,4>$.

2.2. Frequent Pattern Mining

Our work is a case specific extension of frequent pattern mining, first introduced by [2]. The problem of frequent pattern mining could be explained as follows [8]. Let $I=\{i_1,i_2,...,i_m\}$ be a set of m distinct items. Items are ordered by a total order on I. An event (also called itemset) of size l is a non empty set of l items from I^* $(i_1 i_2 ... i_l)$, which is sorted in increasing order. A sequence α of length L is an ordered list of L events $\alpha_1,...,_L$, denoted as $\alpha_1 \rightarrow \alpha_2 \rightarrow ... \rightarrow \alpha_L$, which is ordered based on the timestamp. A sequence database D is composed of sequences, where each sequence has a unique sequence identifier (sid). A sequence $s_a=\alpha_1 \rightarrow \alpha_2 \rightarrow ... \rightarrow \alpha_n$ is contained in (subsequence of) another sequence $s_b=\beta_1\rightarrow\beta_2 \rightarrow ... \rightarrow \beta_m$ if and only if there exist items of integers $1\leq i_1<i_2<...<i_n \leq m$ such that $\alpha_1\subseteq\beta_{i_1}$, $\alpha_2\subseteq\beta_{i_2},...,_n\subseteq\beta_{i_m}$.

We consider medication types in Table I as the list of items and Table II as the sequence database. The patient id pid as the sequence id. An event is a medication episode, which shows a combination of medication(s) that is given to the patient in a period of time. A framework for constructing medication episode from medical history has been introduced in our previous publication. A sequence is a representation of a set of ordered medication episodes based on their timestamps.

Example 2: From Table II, for patient id 2, we have five medication episodes that are the combination of medication type number as follow: $<3,4>$, $<3,5>$, $<1,3,5>$, $<1,4>$, and $<1,4,6>$. The sequence $<3,4> \rightarrow <3> \rightarrow <1,4>$ is contained in patient id 2 sequence. because $<3,4> \subset <3,4>$, $<3> \subset <3,5>$ and $<1,4> \subset <1,4>$. However, the sequence $<1> \rightarrow <4>$ is not contained in $<1,4>$.

The task in frequent pattern mining is to find pattern p which is frequent in a database D if p is contained in at least a certain percentage (support) of sequences of D. The problem in frequent pattern mining is that the number of frequent pattern candidate is

exponential. Therefore, to explore all the possibility could consume a high computation cost. Apriori principle is used to solve the problem. Apriori principle, which is that if a sequence is frequent then the itemsets, which constructed it, must be also frequent. Using the Apriori principle, itemsets that are not frequent and their super sequences will be pruned.

2.3. Singleton Mining

As stated in the previous section, we are interested in investigating the transition events between stable periods. Stable periods represent medications that have proven effective for the patient's conditions. We define a medication pattern as a repeated consecutive of medication found in the SP sequence data set. We consider two types of medication patterns, as follows :

- A singleton pattern is the pattern of a single SP. The singleton pattern may be stated based on the medicine name, medicine type, or medicine label. In the case of diabetes, it may take the form of monotherapy, dual therapy, triple therapy or more.
- n-sequence pattern is a sequence pattern consisting n+1 consecutive singletons.

The support of the pattern is calculated as the ratio of number of patients who exhibit the pattern, at least once in their longitudinal medical history, to the total number of patients, $Sup(Therapy) = \Sigma_{patient\ with\ therapy}|/|\Sigma_{patient}|$.

Example 3: Table II presents the medication type transitions of the stable periods from six patients' medical record. The number inside the bracket denote the medicine type and the arrow represents the transition. For example, the patient with pid 1 has medication transition from dual therapy with medicine types 1 and 3 to dual therapy with medicine types 3 and 5 and then followed by the subsequent transition to dual therapy with medicines types 3 and 6. Hence, from Table II, with a minimum support value of 0.2, we can find four singleton pattern as follows: *<1>*, *<1,4>*, *<3,5>*, *<1,4,6>*. The 1-sequence patterns are as follows : *<1>* → *<1,4>*, and *<1,4>* → *<1,4,6>*.

This pattern definition is different from that used to generate Apriori-based FSP results. In Apriori, it considers that the occurrences of a subset of the singleton support the frequent sequence and the even though the sequence does not in a consecutive manner it will add the cardinality of the pattern.

Example 4: For the data in Table II and a minimum support value of 0.2, Apriori-based FSP mining will indicate *<4,6>* as a frequent pattern with support value of 0.5 and *<1,4>* → *<1,4,6>* as another frequent pattern with a support value of 0.5. By contrast, when singleton mining is applied, *<4,6>* is not considered as a frequent pattern because it is not a full itemset, where as for *<1,4>* → *<1,4,6>* pattern, the calculated support value is only 0.33 because the patient with pid 6 is not counted as supporting the pattern support.

2.4. Association Rules

The final result of frequent pattern mining is attaining association rules. The rules is defined as an implication of the form:

$$X \Rightarrow Y,\ where\ X, Y \subset Items \qquad (1)$$

From the definition, we could separate the rules into two parts, the left side is the antecedent (if- part) and the right side is the consequent (then- part).

Example 5. If we use minimum support value of 0.2, then we will have association rules *<4>* ⇒ *<6>* and *<1,4>* ⇒ *<1,4,6>* from the frequent sequence patterns when we applied Apriori. For the case of singleton, we will have rules *<3>* ⇒ *<5>* and *<1,4>* ⇒ *<1,4,6>*.

2.5. Data Visualization Using Graph

Definitions used in a graph theory is given as the following [8]: A graph is a set *V* together with a relation on *V*. A graph *G = (V,E)* is a pair of sets, *V* is a set of nodes (vertices), *E* is a set of edges (arcs or links). An edge *e (u,v)*, with, *e* ∈ *E* and *u,v* ∈ *V*, is a pair of vertices. If the relation on *V* induced by *E* is symmetric; we call such a graph undirected. If the pair of vertices in an edge is ordered, *G* is denotes as directed graph or digraph. Direction is denoted by saying, with respect to a node, that an edge is incoming or outgoing. A graph is weighted if each of its edges is associated with a real number. An unweighted graph is equivalent to a weighted graph whose edges all have a weight of 1. A graph is complete if there exists an edge for every pair of vertices. If it has n vertices, then a complete graph has *n(n-1)/2* edges. A loop is an edge with *u=v*. A path is a list of successively adjacent, distinct edges. Let $<e_1,..., e_k>$ be a sequence of edges in a graph. This sequence is called a path if there are vertices $<v_1,...,v_k>$ such that $e_1=v_i\text{-}1,v_o$ for *i=2,...,k*. A path is cyclic if a node appears more than once in its corresponding list of edges. A graph is cyclic if any path in the graph is cyclic and acyclic if there are no cyclic path in the graph. A tree is a graph in which any two nodes are connected by exactly one path. Trees are thus acyclic connected graphs. trees may be directed or undirected. A tree with one node labeled root is a rooted tree.

3. Method

In this section, we describe our method to visualize association rules by developing a medication trajectory graph.

3.1. Visualizing 1-sequence patterns into a directed graph

The requirements of the medication trajectory graph are as the following:

1. The graph should be an acyclic graph (i.e, a rooted tree graph with left to right direction).
2. Nodes and edges are frequent patterns in the form of singleton and 1-sequence pattern produced by singleton mining
3. A node is a singleton, which represents combination of medication and an edge is a sequence of adjacent singletons, which represents a medication transition event. The node and edge are associated with the support of the pattern.
4. The root should be a monotherapy and then, propagate into dual therapy, triple therapy and so on.
5. Nodes' existence in later level (consequent) are dependent to nodes from the previous level (antecedent). Thus nodes in each level have parent(s) or incoming edges (except root) but the nodes may have no children or outcoming edges.
6. The graph should not contain loop.

An example of medication trajectory model is shown in Fig. 1. Fig. 1 shows a three levels of tree graph. the first level is the root (monotherapy), the second level should be dual therapies, and the third level should be triple therapies. And as demonstrated by Fig. 1, there are no loops in the graph (i.e., incoming edge from node in the later level).

To enable dynamic visualization, we develop the graph using PHP, which is a general-purpose scripting language [9], and Vis.js, which is a javascript library for dynamic, browser based visualization [10]. The algorithm is shown in Algorithm 1. The weight retrieval is not described in the algorithm. However, it can be easily retrieve along with the singleton and 1-sequence patterns

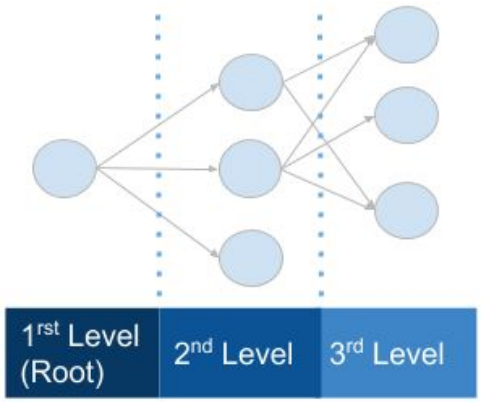

Fig. 1. An example of medication trajectory model

retrieval respectively are pushed into *N* and *E* hash respectively to the nodes and edges.

Algorithm 1 Build Medication Trajectory Graph

Input: a set of singletons *S* and 1-sequence patterns *T* from the result set of singleton mining;
Output: trajectory model;
begin
  select a monotherapy *root* from *S* and maximum combination number *maxLevel* from *T;*
  initialize previous level node *prevAnt = root;*
  initialize nodes *N* and edges *E* as hashes;
  for *level*=2 **until** *maxLevel*
    select 1-sequence pattern(s) *e,* destination node(s) *con(s)* from *T* having source node(s) *ant(s)* equal to *prevAnt(s)* and *con(s)* is not the same with the *prevAnt(s)* from *T;*
    if *con* is not in N **then** push con into N;
    push *e* into *E;*
  end for
  build directed graph with *N*, *E* data;
end

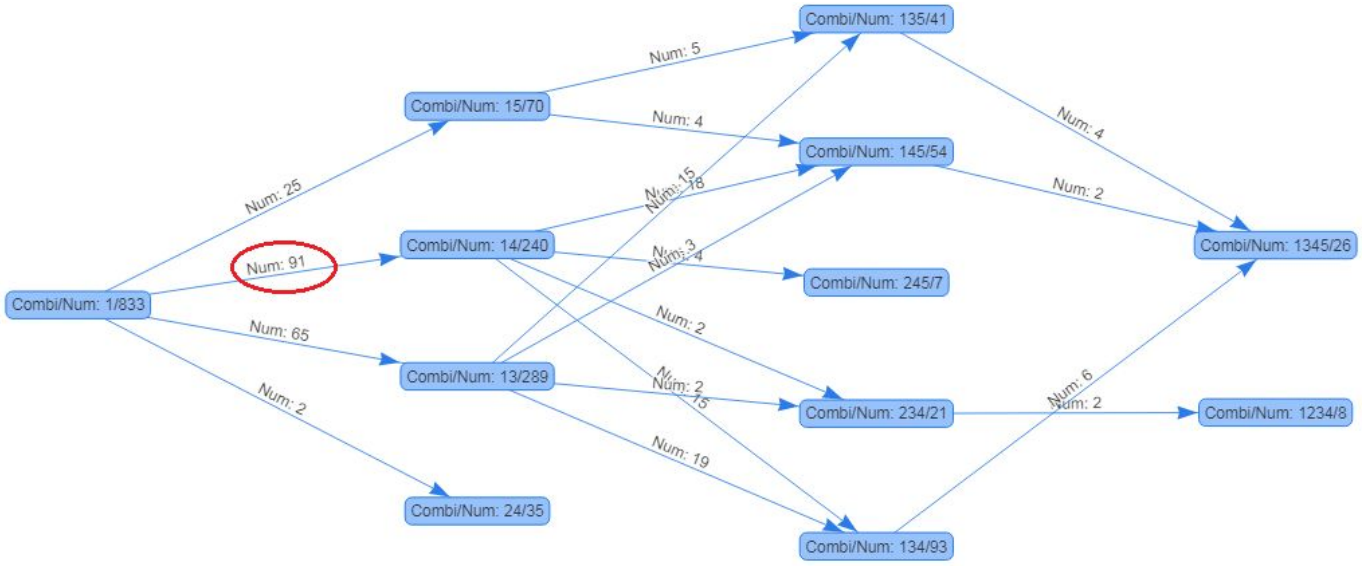

Fig. 2. Medication trajectory graph for prior 2010

4. Visualization and Discussion

We use stable period sequences of Type 2 diabetes patients, which are identified from medication episodes construction framework on medical history provided by Kyoto University Hospital along with the approval from the Ethics Review Board of The Medical School of Kyoto University.

In this experiment, we conduct visualizations for dataset from 2000 until 2009. From this dataset, we have stable period sequences from 1781 patients. In addition, the nodes and edges are patterns having minimum support of 0.001.

The result of the visualisation prior 2010 is shown in Fig. 2. The label inside the nodes shows the medication combination and the number of patient having the pattern. For example from Fig. 2 in root node, the medication combination is medication type 1 that is Sulfonylurea (SU) and there are 833 patients prescribed with SU. The edge from node 1 toward node 15 is a medication transition from medication type 1 to combination of medication type 1 and 5; and there are 25 patients having this transition.

Using this visualization, the domain expert is able to identify quickly a particular interesting rule that is the edge from 1 to 14, which is a transition event from medication type 1 to medication combination of type 1 and 4 (Biguanides). This transition pattern is occured in 91 patients.

5. Conclusion

We describe a method for visualizing the association rules from 1-sequence pattern to form a medication trajectory graph. The results show that the method is able to help domain expert to identify interesting rule.

6. Acknowledgement

The first author is supported by Indonesian Endowment Fund for Education (LPDP). This work was supported by JSPS KAKENHI Grant Number JP15H02705.

References

[1] Treatment Guide for Diabetes 2012-2013, Japan Diabetes Society.

[2] Agrawal, Rakesh, and Ramakrishnan Srikant. "Mining sequential patterns." Data Engineering, 1995. Proceedings of the Eleventh International Conference on. IEEE, 1995.

[3] Khotimah, Purnomo Husnul, et al. "Comparing frequent patterns: A study case of Apriori and singleton implementations in a diabetes type 2 data set." Computer, Control, Informatics and its Applications (IC3INA), 2016 International Conference on. IEEE, 2016.

[4] Liu, Rong, Jianzhong Zhou, and Ming Liu. "Graph-based semi-supervised learning algorithm for web page classification." Intelligent Systems Design and Applications, 2006. ISDA'06. Sixth International Conference on. Vol. 2. IEEE, 2006.

[5] LeBlanc, Michael, Joth Jacobson, and John Crowley. "Partitioning and peeling for constructing prognostic groups." Statistical methods in medical research 11.3 (2002): 247-274.

[6] Rafi, Muhammad, Farnaz Amin, and Mohammad Shahid Shaikh. "Document clustering using graph based document representation with constraints." arXiv preprint arXiv:1412.1888 (2014)..

[7] National Health Priority Action Council (NHPAC) (2006), National Chronic Disease Strategy, Australian Government of helath and Ageing, Canberra.

[8] Chen, Chun-houh, Wolfgang Karl Härdle, and Antony Unwin, eds. “Handbook of data visualization.” Springer Science & Business Media, 2007.

[9] PHP Group. “PHP.”, available online at http://www.php.net/

[10] Almende B.V..”Vis.js.”, available online at http://visjs.org/