Association Rule Algorithm Sequential Pattern Discovery using Equivalent Classes (SPADE) to Analyze the Genesis Pattern of Landslides in Indonesia

Muhammad Muhajir*1,2, Berky Rian Efanna*2

* Statistics, Faculty of Mathematics and Natural Sciences Islamic University of Indonesia
1 muhammad.muhajir.stat89@gmail.com*; 2 berkyeki@gmail.com

I. Introduction

Indonesia is a disaster-prone country because it is located between the confluence of three major plates that active in the world like Eurasian plate, Indo-Australian plate and Pacific plate. According to the Law No. 24 of 2007 on Disaster Management, disaster is an event or series of events that threaten and disrupt the lives and livelihood caused by natural factors or factors of non-natural or human factors that lead to the emergence of human lives, environmental damage, loss of property, and the psychological impact.

According to BNPB the highest number of disaster events is landslides as much as 402 events occurs until August 2015. Landslide is one of movement of soil, rock, soil creep, and rock debris that occurred once the move to the slopes. It is caused by steep slopes, high rainfall, deforestation, mining activities, and erosion. The impacts of the landslide are loss of property, damage to facilities such as homes and buildings, casualties, psychological trauma, disrupted economic and environmental damage. Based on the impacts of landslide, mitigation required to take early precautions are to know how the pattern of association between the sequence of events landslides and to know how the associative relationship pattern of earthquakes. Based on the impacts, the results of this research is associative relationship pattern is obtained from data flood that occurs in Indonesia, namely in case of heavy rain will occur labile soil structure to support the value of 0.37, confidence level of 41% and the power of formed ruled is 1.02.

Copyright © 2015 International Journal of Advances in Intelligent Informatics. All rights reserved.

Keywords: SPADE, Landslide, Association Rule, Algorithm.
II. Related Works

Applied research related to disaster especially landslide has been investigated by several researchers. First, analyzing the Land use change and the landslide characteristics for community-based disaster mitigation. The results show that a change in vegetation cover results in a modified landslide area and frequency and changed land use areas have higher landslide ratios than no changed. Land use management and community-based disaster prevention are needed in mountainous areas of Taiwan for hazard mitigation [4]. Second, analyzing the Landslide damage and disaster management system in Nepal. The results show that the landslide in Nepal was mainly caused by the combine effect of high rainfall, a steep slope and unconsolidated rock at the bed. The debris mass flowed along with the flood and caused damage downstream of the watershed. The existing landslide disaster management system in Nepal is weak so the disaster management system in Nepal must be considered as a part of rural development [5]. There is method to analyzing a framework for regional association rule mining and scoping in spatial datasets can be applied also in landslide case. The results of this research are spatial risk pattern and risk zones of arsenic in the Texas water supply were obtained [6]. However, these studies used Mining Conjunctive Sequential Pattern. The results from this paper is the new introduced patterns have high potential for real life applications like landslide case [7].

III. Basic Theory

A. Association Rule

Association rules is one of the main techniques in data mining and the most commonly used in finding a pattern or patterns from a data set [8]. Support is a measure that indicates the degree of dominance of an item or the entire item set transaction [9]. Support in this study is the probability of the sequence of events in a single incident of landslide is interconnected with the overall incidence of others landslides [10]. Thus, the value of an item support calculated as (1).

$$Support\ (X) = P(X) = \frac{n(X)}{n(S)}$$ (1)

Where $P(X)$ is a probability of event $X$, $n(X)$ is a number of event $X$ in transaction, and $n(S)$ is the number of transactions on database $S$. Confidence is a strong relationship between items in association rules. In this research, confidence is defined as the probability of occurrence of certain items (the chronology of the landslide) in a single event (interconnected) and one of the chronology is certainly due to several causes of the landslide. Thus, the value of a combination of items confidence calculated as (2).

$$Confidence\ (X => Y) = P(Y/X) = \frac{P(X \cup Y)}{P(X)}$$ (2)

Where $P(Y/X)$ is a conditional probability of occurrence of Y when X events occurred, $P(X \cup Y)$ is a probability of occurrence of X and Y simultaneously, and $P(X)$ is a probability of occurrence X. Besides to these two parameters, one of the better ways to determine the strength of an association rule is to look at the value of the lift ratio. Lift ratios indicate the power level of the rule on random events of the antecedent (X) and consequence (Y) based on the each support expressed in equation of (3).

$$Lift\ ratio\ (X => Y) = \frac{P(X \cup Y)}{P(X)P(Y)}$$ (3)

Where $P(Y/X)$ is the probability of occurrence of events X and Y simultaneously, $P(X)$ is a probability of occurrence X, and $P(Y)$ is a probability of occurrence Y.
B. Sequence Pattern Mining

Sequential pattern mining used for data that has a sequence, the data can be a sequence of transactions. Sequential pattern mining first introduced by Agrawal and Srikant. Sequential pattern mining process can be described as follows, for example given a number of sequences, each sequence consisting of a series of elements and each element of support. Excavation sequential pattern is all of subsequence search repeated, subsequence that has the bigger frequency of occurrence than the minimum-support consists of a number of items, and given the minimum value 11. To settle this sequential problem can be done by several methods. One of the methods is SPADE (Sequential Pattern Discovery Using Equivalence Classes).

C. SPADE (Sequential Pattern Discovery Using Equivalence Classes)

SPADE algorithm (Sequential Pattern Discovery using Equivalence classes = Invention of data sequence pattern using the same class) is a new algorithm for rapid discovery of data sequence pattern 11. The definition of the class is a collection of objects that have the same attributes or parameters, while the frequency is the number of times data has the same value. The problem of data mining sequence patterns can be expressed as follows: I = \{i_1, i_2, \ldots, i_n\} an object consisting of a set of alphabet. While an event is a collection of actions that have orders to do. Sequence is a list of events. An event is denoted as (i_1, i_2, \ldots, i_k), where j is the object. If there is an α which is a sequence of objects that can be denoted as follows (α_1 → α_2 → \cdots → α_q), where α is an incident. A sequence with k objects denoted by k =Σ|α|=1| then this means that k is a k-order (k-sequence).

SPADE algorithm steps in finding frequent sequence and then to determine the rules of the frequent sequence are as follows 12.

1. Determine frequent 1-sequence
   - Do the scan for each item set in a sequence database.
   - Save the id-list for each item set (sid and eid pair).
   - Then scan the id-list from each id-list, each encountered sid that did not exist before, then value of the support is added.
   - Sequence that entered in frequent 1-sequence is the support that have value of more than \textit{min\_sup}.

2. Determine frequent 2-sequence
   - The data that used is data of frequent 1-sequence.
   - Combine each frequent 1-sequence with all other frequent 1-sequences. For example, if 1-sequence \textit{A} merged with 1-sequence \textit{B}, then the possibility of two sequences that occurs is \textit{A,B} where \textit{A} and \textit{B} appear together in the transaction, \textit{A → B} where item \textit{B} appear after item \textit{A} and \textit{B → A} where item \textit{B} appears after item \textit{A}.
   - Check the id-list whether the id-list is have the equal sid for every merger of frequent 1-sequence, if equal, then check the eid of 1-sequence \textit{A} is equal or less than or more than eid 1-sequence \textit{B}.
   - If equal, then id-list is included in the 2-sequence A,B. If \textit{eid B} is more than \textit{A}, then the id-list is included in the 2-sequence \textit{A → B}. If \textit{eid A} is more than \textit{B}, then the id-list is included in the 2-sequence \textit{B → A}.
   - Then, as in the frequent 1-sequences, add the support for each sid that did not exist before.
   - From the 2-sequence check the support value whether the support is more than \textit{min\_sup} or not. If the support value is eligible, then it is entered in frequent 2 sequence.

3. Determine frequent k-sequence
   - After determined the frequent 2-sequence, do the same process to seek the next frequent sequence, which is to determine frequent k-sequence. To determine a frequent k-sequence is performed to join the frequent (k-1) sequences that have the same prefix. For example, to determine the 3-sequence combine the frequent sequence of 2-sequences that have the same prefix, to determine the 4-sequence combine the frequent sequence of 3-sequences that have the same prefix, and so on. To determine prefix frequent (k-1) sequence remove the last item of the sequence. For example, if there is a 4-sequence \textit{A → B → C → D}, then the prefix is \textit{A → B → C}. For each of this merger there are 3 possible outcomes:
     - If \textit{A, B} are combined with \textit{A,C}, then the possible result only \textit{A, B, C}.
• If A,B are combined with A → C, then the possible result only A,B → C.
• If A → B → C combined with A, then there are 3 possible outcomes: A → B, C., A → B → C and A → C → B.

From each of these possibilities, check the support value. Whether it meets the min_sup or not. If yes then the sequence was included in the frequent k-sequence. Frequent sequence’s searching is terminated if there is no frequent (k-1) sequences that could be join or there is no frequent k-sequence that found anymore.

4. Establishment of Rule
• After all frequent sequences are found, determined the rule of these sequences.
• 1-sequences are not used to establish the rule because it is only consists of one item.
• To 2-sequence which is antecedent is the first item and the consequent is the second item. Examples for sequence A → B then established rule is A ⇒ B. As for the sequence which is longer than 2 or k-sequence, the last item is used as consequent, while antecedent are all the items before the last item.
• For example, to 4-sequence A → B → C → D, then the established rule is A → B → C ⇒ D. Calculated the confidence value for each rule. If it meets the limits of min_conf rule, then the rule is accepted.

IV. Results and Discussion
The population in this research is the occurrence landslides’ data in Indonesia and the sample is the chronology of landslides events period August 2011 until June 2015. The type of data that will be used is secondary data obtained from the website of Indonesian National Board for Disaster Management. Variables of this research is chronologic. Chronologic is the sequence of events that occurred in landslide. This research using Sequential Pattern Algorithm Discovery using Equivalent Classes (SPADE). There are four step SPADE algorithm as follows:

A. Selection Data
This research using landslide data from Indonesian National Board for Disaster, as examples of the data is shown in Table 1.

<table>
<thead>
<tr>
<th>No</th>
<th>Date</th>
<th>Location</th>
<th>Victim</th>
<th>Loss</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>29/06/2015</td>
<td>Subdoladi, Bantul District, Yogyakarta, Indonesia</td>
<td>None</td>
<td>1 Unit broken bridges and several wells affected by landslide material</td>
<td>Chronologic: Caused by dredging the land in the cliff area Effort: Monitoring By BPBD District. Bantul, Indonesia</td>
</tr>
<tr>
<td>2</td>
<td>06/06/2015</td>
<td>Keroyo, Mekarsari, Sajira,Lebak, Banten, Indonesia</td>
<td>1 Victim (Mr. Makmun, 50 years old)</td>
<td>None</td>
<td>Chronologic: At the time of the landslide, the victim were digging kalimaya stone, rocks falls on the head of the victim and the victim’s head was leaking. Effort: The victim was taken to the nearest health center.</td>
</tr>
</tbody>
</table>

Based on these data it appears that the data landslides has many attributes such as date, location, victim, Losses, and Description. However, not all of them will be used in this research so that the data do preprocessing to acquire the attributes used in the study.

B. Cleaning Data
This phase will clean up the data that is not needed to reduce data errors and duplication of data. For example, removing the attributes such as Date, Location, Victim, Losses, because that attributes are not used to establishment the rule. The data that used is an example of the result of the cleaning process is shown in Table 2.
### Table 2. Cleaning Data

<table>
<thead>
<tr>
<th>No</th>
<th>Abbreviation</th>
<th>Chronological</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PT</td>
<td>Dredging Soil</td>
</tr>
<tr>
<td>2</td>
<td>TL</td>
<td>Landslides</td>
</tr>
</tbody>
</table>

### C. Transformation Data

After data with attribute (numbers of events and chronologic) was obtained as presented in Table 2, the next phase is transformation of data by making a co-occurrence table. Co-occurrence table illustrates the strong collection of investigated landslide occurrences and also the chronology of landslides’ events. As for Association Sequential Pattern Mining with SPADE algorithm, first the data is transformed into data format vertical, then the database sequences to shape a set sequence with format [itemset: (sequence_ID, EventID)]. In other words, for each itemset will be stored as sequence identifier and event identifier corresponding. Event identifier is useful as a timestamp of the itemset. A pair (sequence_ID, EVENT_ID) for each itemset shape ID_LIST of itemset. Some examples can be seen in Table 3.

### Table 3. Transformation Data

<table>
<thead>
<tr>
<th>No</th>
<th>Items</th>
<th>Sequence ID</th>
<th>Event ID</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{PT}</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>{TL}</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>{HD}</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>{TL}</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>{PB}</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>{TL}</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>{GT}</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>{TL}</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>{HL}</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>{TL}</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

After the data is in accordance with the format SPADE algorithm obtained, the next analysis can be done. The next analysis is to look for the sequential pattern.

### D. Sequential Pattern Mining with SPADE Algorithm

Sequential Pattern Mining with SPADE algorithm with minimum limit support 0.4 and the minimum limit confidence 0.01 used to know the association pattern between the chronologies of the landslides’ events. Table 4 show the result of the analysis using statistical software R 3.2.2.

### Table 4. Association with SPADE Algorithm

<table>
<thead>
<tr>
<th>No</th>
<th>Assosiations</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{ Heavy Rain } =&gt; { Landslides }</td>
<td>0.89</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>{ Labile Soil Structure } =&gt; { Landslides }</td>
<td>0.40</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>{ Heavy Rain, Labile Soil Structure } =&gt; { Landslides }</td>
<td>0.37</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>{ Heavy Rain } =&gt; { Labile Soil Structure }</td>
<td>0.37</td>
<td>0.41</td>
<td>1.02</td>
</tr>
</tbody>
</table>

Based on the defined limits that the minimum support 0.3 and the minimum confidence 0.1 obtained four association rules are formed and the information that obtained from these rules are:

a. Rule \{Heavy Rain\} => \{Landslides\}

Rules with support value of 0.89, confidence and lift ratio of 1 to 1. The meaning of support value of 0.89 is 89% or 552 landslides of the whole landslides (620 landslides) caused by heavy rainfall. Confidence value of 1 means that in case of heavy rain will causes landslides with a confidence level of 100%. While lift ratio value of 1 indicates how strong the rule or rules formed of sequential pattern mining algorithms. Lift ratio value ranged from 0 to infinity. If the
lift ratio value equals to 1, then the rule \{Heavy Rain\} => \{Landslides\} often occurred together but independently.

b. Rule \{ Labile Soil Structure\} => \{Landslides\}

Rules with support value of 0.40, confidence and lift ratio of 1 to 1. The meaning of support value of 0.40 is 40% or 248 landslides of the whole landslides studied (620 landslides) caused by unstable soil structure. Confidence value of 1 means that if the soil structure is unstable, then it will causes landslides with confidence level of 100%. While lift ratio value of 1 indicates how strong the rule or rules formed of sequential pattern mining algorithms. Lift ratio value ranged from 0 to infinity. If the lift ratio value equals to 1, then the rule \{Labile Soil Structure\} => \{Landslides\} often occurred together but independently.

c. Rule \{Heavy Rain, Labile Soil Structure\} => \{Landslides\}

Rules with support value of 0.37, confidence and lift ratio of 1 to 1. The meaning of support value of 0.37 is 37% or 230 landslides of the whole landslides studied (620 landslides) caused by heavy rains and unstable soil structure. Confidence value of 1 means that heavy rains and unlabile soil structure will causes landslides with confidence level of 100%. While lift ratio value of 1 indicates how strong the rule or rules formed of sequential pattern mining algorithms. Lift ratio value ranged from 0 to infinity. If the lift ratio value equals 1 then the rule often occur together but independently. If the lift ratio value of more than 1, the rules will be recommended because the antecedent has a positive influence on the consequent. Rules with lift ratio value of more than 1 can be interpreted as a powerful rules. If lift ratio value of 1.02, then the rule \{Heavy Rain\} => \{Labile Soil Structure\} is recommended.

d. Rule \{Heavy Rain\} => \{ Labile Soil Structure \}

Rules with support value of 0.37, confidence and lift ratio of 1 to 1.02. The meaning of support value of 0.37 is 37% or 230 landslides of the whole landslides studied (620 landslides) caused by heavy rainfall. Confidence value of 0.41 means that in case of heavy rain will causes unstable soil structure with confidence level of 41%. While lift ratio of 1.02 indicates how strong the rule or rules formed of sequential pattern mining algorithms. Lift ratio value ranged from 0 to infinity. If the lift ratio value equals 1 then the rule often occur together but independently. If the lift ratio value of more than 1, the rules will be recommended because the antecedent has a positive influence on the consequent. Rules with lift ratio value of more than 1 can be interpreted as a powerful rules. If lift ratio value of 1.02, then the rule \{Heavy Rain\} => \{Labile Soil Structure\} is recommended.

V. Conclusion

Based on the results of the analysis can be concluded that obtained four rules using minimum support value of 0.3 and minimum confidence value of 0.1 as follows:

1. Rules \{Heavy Rain\} => \{Landslide\}
2. Rules \{ Labile Soil Structure \} => \{Landslide\}
3. Rules \{Heavy Rain, Labile Soil Structure\} => \{Landslide\}
4. Rules \{Heavy Rain\} => \{ Labile Soil Structure \}

The most recommended rules is rules number four, rules \{Heavy Rain\} => \{Structural Soil labile\} with support value at 0.37, confidence value and lift ratio value of 0.41 to 1.02. Support of 0.37, confidence and lift ratio of 1 to 1.02. The meaning of support value of 0.37 is 37% or 230 landslides of the whole landslides studied (620 landslides) caused by heavy rainfall. Confidence value of 0.41 means that in case of heavy rain will causes unstable soil structure with confidence level of 41%. While lift ratio of 1.02 indicates how strong the rule or rules formed of sequential pattern mining algorithms.

References


