InfoMiner+: Mining Partial Periodic Patterns with Gap Penalties

Jiong Yang
UIUC
jioyang@cs.uiuc.edu

Wei Wang
UNC-Chapel Hill
weiwang@cs.unc.edu

Philip S. Yu
IBM T. J. Watson Research Center
psyu@us.ibm.com

Abstract

In this paper, we focus on mining periodic patterns allowing some degree of imperfection in the form of random replacement from a perfect periodic pattern. Information gain was proposed to identify patterns with events of vastly different occurrence frequencies and adjust for the deviation from a pattern. However, it does not take any penalty if there exists some gap between the pattern occurrences. In many applications, e.g., bio-informatics, it is important to identify subsequences that a pattern repeats perfectly (or near perfectly). As a solution, we extend the information gain measure to include a penalty for gaps between pattern occurrences. We call this measure as generalized information gain. Furthermore, we want to find subsequence S' such that for a pattern P, the generalized information gain of P in S' is high. This is particularly useful in locating repeats in DNA sequences. In this paper, we developed an effective mining algorithm, InfoMiner+, to simultaneously mine significant patterns and the associated subsequences.

1 Introduction

Periodic pattern discovery is an important problem in mining time series data and has wide applications. A periodic pattern is a list of ordered events, which repeat itself in the event sequence. It is useful in characterizing the cyclic behavior of the time series. In practice, not every portion in the time series may contribute to the periodicity. For example, a company's stock may often gain a couple of points at the beginning of each trading session but it may not have much regularity at later time. This kind of looser periodicity is often referred to as *partial periodicity*. Moreover, due to some random noise, a pattern may not always be repeated perfectly. In turn, the event sequence can be viewed as a series of perfect pattern repetitions with a few *random replacements*[1]. If the amount of "replacement" is below some reasonable threshold, we may regard that the pattern still exhibits in the event sequence.

As a newly developed research area, most previous work on mining time series data addresses the issue by creating a mapping to the association rule mining technique and therefore uses the support and confidence as the metrics to identify the significant patterns from the rest. Most association rule mining algorithms favor frequently occurred event(s) due to the nature of the problem. However, patterns involving infrequent events may also be as significant as (or even more significant than) frequent events in an event sequence. This issue becomes more critical when different events occur at divergent frequencies.

Information gain is introduced in [3, 4, 5, 6] to measure the importance/significance of the occurrence of a pattern. The information gain of an occurrence of a rare event is high while the information gain of the occurrence of a frequent event is low. Thus, we are able to find the statistically significant patterns with the information gain threshold. However, the major limitation of this model is that it does not take into account where the occurrences are in the sequence. Let's take a look at two sequences: $S_1 = a_1, a_2, a_3, a_3, a_2, a_3, a_1, a_2, a_4, a_5, a_1, a_2$ and $S_2 = a_1, a_2, a_1, a_2, a_1, a_2, a_3, a_3, a_2, a_3, a_4, a_5$. The elements in the two sequences are identical. The only difference is the order of the events. The pattern (a_1, a_2) repeats perfectly in the first half of S_2 while it scatters in S_1. The two patterns have the same information gain in the two sequences. In some applications (e.g., repeats discovery in bio-informatics domain), a series of consecutive repeats are considered more significant than the scattered ones. That is, there is some "penalty" associated with the gap between pattern repeats [2]. As a result, we need a measurement that can make such a distinction. In this paper, we introduce the measure of generalized information gain (GIG) to capture the significance of a pattern in a sequence/subsequence. The occurrence of a pattern will be given a position GIG while a mis-occurrence (or a gap) will generate a negative GIG. The overall generalized information gain will be the aggregate of all occurrences and mis-occurrences of the pattern in a sequence/subsequence.

Since the characteristics of a sequence may change over time, many patterns may only be valid for a period of time. The degree of significance (i.e., generalized information gain) of a pattern may be diluted if we only consider the entire event sequence. In addition, a user may be interested not only in a significant pattern, but also where/when the pattern is significant as well. In a DNA sequence, researchers are also interested in which region a pattern is significant. The identification of significant pattern in a subsequence is of great importance in many applications. However, most previous research

[1] Or equivalently, the event sequence would become a series of perfect repetitions of some pattern after a few replacements (of events).

focused on discovering patterns that are significant over the entire data sequence[6]. In this paper, we relax the model by allowing a pattern to be significant during a contiguous portion of the entire sequence. A user is asked to specify the minimum GIG a significant pattern must carry over a subsequence of data. Upon satisfying this requirement, the subsequence(s) that maximizes the GIG of a pattern will be identified. In the previous example, the pattern (a_1, a_2) is very significant in the first half of S_2, but may not be significant over the entire sequence.

Although the generalized information gain is a more meaningful metric for the problems addressed previously, it does not preserve the *downward closure* property (as the *support* does). For example, the pattern $(a_1, a_2, *)$ may have sufficient GIG while both $(a_1, *, *)$ and $(*, a_2, *)$ do not[2]. We can not take advantages of the standard pruning technique developed for mining association rules. The observation that the *triangle inequality*[3] is still preserved by the generalized information gain motivates us to devise a threefold algorithm as the core of our pattern discovery tool, InfoMiner+.

1. First, the significant patterns involving one event are discovered.
2. Next, candidate patterns involving multiple events are generated based on the *triangle inequality*.
3. All candidate patterns are validated and for each pattern which is significant, the corresponding subsequence containing the pattern is also identified.

The remainder of this paper is organized as follows. The problem is formulated in Section 2 while our approach is presented in Section 3. We conclude this paper in Section 4.

2 Problem Formulation

Let $\Im = \{a_1, a_2, \ldots, \}$ be a set of events and D be a sequence of events in $\Im$. We now introduce some notations that we will use for the remainder of the paper. As mentioned previously, we want to discover the patterns P such that there exists some subsequence D' of D, where P is significant in D'.

Definition 2.1 *A* **pattern** *with* **period** l *is an array of* l *events* $(p_1, p_2, \ldots, p_l)$, *each of which is either an event in* $\Im$ *or* *, *i.e.,* $p_j \in \Im \cup * \ (1 \leq j \leq l)$. *We say that the* j*th position is* **instantiated** *if* $p_j \in \Im$. *For any two patterns* $P = (p_1, p_2, \ldots, p_l)$ *and* $P' = (p'_1, p'_2, \ldots, p'_l)$ *of the same period* l, P *is a* **superpattern** *of* P' *(*P' *is a* **subpattern** *of* P*) if* $p'_j = *$ *or* $p_j = p'_j$, *for all* $1 \leq j \leq l$.

[2] We will explain it in more detail later in this paper.

[3] For example, the GIG of $(a_1, a_2, *)$ can not exceed the summation of that of $(a_1, *, *)$ and $(*, a_2, *)$.

Note that an event can appear at multiple positions in a pattern. For example, $(a_1, a_4, *, *, a_4)$ is a pattern of period 5 and its first, second and fifth positions are instantiated. It is also a superpattern of $(a_1, *, *, *, a_4)$.

We found that the *information* measure [1] which is widely studied and used in information theory field fulfills the first two properties. It is hence used to measure the amount of positive/negative influence to the overall significance of a pattern caused by each occurrence/absence. Based on it, we devise a novel measurement, *generalized information gain*, to quantify the significance of a pattern within a subsequence of events. The following is the definition of the information and the generalized information gain.

Definition 2.2 *For an event* $a_j \in \Im$ *and a sequence* D *of* N *events, let* $Prob(a_j)$ *be the expected probability that* a_j *occurs at any given position in* D[4]. *Then the* **information** *of* a_j *with respect to* D *is defined as* $I(a_j) = \log \frac{1}{Prob(a_j)} = -\log Prob(a_j)$. *The information of the "eternal" event * is always 0*[5].

In the information theory community, information is originally used to quantify the *a priori* uncertainty that an event will occur or the "surprise" if the event occurs; and is defined as a continuous function of the probability that an event occurs. When the probability of an event occurrence approaches 1, the information of that event approaches 0. This enables us to handle the eternal event seamlessly. In addition, information is additive for independent events. This does not only provide a solid theoretical foundation, but also offer computational efficiency.

Definition 2.3
The **information** *of a pattern* $P = (p_1, p_2, \ldots, p_l)$ *is the summation of the information carried by each individual position, i.e.,* $I(P) = \sum_{1 \leq j \leq l} I(p_j)$.

Definition 2.4 *Given a pattern* $P = (p_1, p_2, \ldots, p_l)$ *with period* l *and a sequence of* l *events* $D' = d_1, d_2, \ldots, d_l$, *we say that* D' *is* **in compliance with** P *at position* j $(1 \leq j \leq l)$ *iff either* $p_j = *$ *or* $p_j = d_j$ *holds.*

For example, the sequence a_1, a_1, a_2, a_3 is in compliance with the pattern $(a_1, a_2, a_3, *)$ at positions 1 and 4.

Definition 2.5 *Given a pattern* $P = (p_1, p_2, \ldots, p_l)$ *with period* l *and a sequence of* l *events* $D' = d_1, d_2, \ldots, d_l$, *we say that* P **matches** D' *(or* D' **supports** P*), iff* D' *is in compliance with* P *at every position* $j (1 \leq j \leq l)$.

[4] For the sake of simplicity of exploration, we assume that, without additional qualification, a_j occurs equally likely at any position with probability $Prob(a_j)$. All results presented in this paper can be modified to apply to a more general scenario.

[5] Another way of looking at it is that $Prob(*) = 1$ at any time.

For instance, the sequence a_1, a_4, a_2, a_3, a_4 supports the pattern $(a_1, a_4, *, *, a_4)$ while the sequence a_1, a_4, a_2, a_3, a_6 does not support it since the sequence is not in compliance with the pattern on the last position.

Definition 2.6 *Given a pattern P with period l and a sequence D of $N(N \geq l)$ events: $d_1, d_2, \ldots, d_N$, the* **support** *of P within D is the number of subsequences $d_{l\times j+1}, d_{l\times j+2}, \ldots, d_{l\times j+l}$ that support P.*

Intuitively, the event sequence can be viewed as a list of segments, each of which consists of l contiguous events. There would be $\lfloor N/l \rfloor$ full segments, among which the segment that P matches will count for the support of P.

Definition 2.7 *Given a pattern $P = (p_1, p_2, \ldots, p_l)$ with period l and a sequence of l events $D' = d_1, d_2, \ldots, d_l$, the* **information loss** *of D' on position j with respect to P is the information of the event p_j iff D' is not in compliance with P at position j and there is no information loss otherwise. The overall information loss of D' with respect to P is the summation of the information loss of each position.*

Definition 2.8 *Given a pattern P with period l and a sequence D of $N(N \geq l)$ events: $d_1, d_2, \ldots, d_N$, the* **information loss** *of D with respect to P is the summation of the information loss of each segment $d_{l\times j+1}, d_{l\times j+2}, \ldots, d_{l\times j+l}$ with respect to P. The* **generalized information gain** *of D with respect to P is defined as $I(P) \times (S_D(P) - 1) - L_D(P)$ where $I(P)$, $S_D(P)$, and $L_D(P)$ are the information of P, the support of P within D, and the information loss of D with respect to P, respectively.*

In a subsequence, the first match of a pattern is viewed as an example, and only subsequent matches contribute to the generalized information gain[6].

Definition 2.9 *Given a pattern P, a sequence D and a generalized information gain threshold g, if there exists a subsequence D' of D so that the generalized information gain of D' with respect to P is at least g, then P is a* **valid pattern**.

Theoretically, the period of a valid pattern could be arbitrary, i.e., as long as the event sequence. In reality, a user can specify an upperbound of period length according to his/her domain knowledge. As a result, we use L_{max} to denote the maximum period allowed for a pattern. However, L_{max} can be arbitrarily large, e.g., ranging to several thousands. Now we can rephrase our problem model by employing the generalized information gain metric. For a sequence of events D, an information gain threshold g, and a period bound L_{max}, we want to discovery all valid patterns P whose period is less than L_{max}. In addition, for each valid pattern P, we want to find the subsequence which maximizes the generalized information gain of P. In the remainder of this section, we give some more definitions which enable us to present our approach and communicate to readers more effectively.

[6] Since we aim at mining periodic patterns, only repeated occurrences of a pattern are used to accumulate the generalized information gain.

Definition 2.10 *For any two patterns $P = (p_1, p_2, \ldots, p_l)$ and $P' = (p'_1, p'_2, \ldots, p'_l)$ of the same period l, P and P' are* **complementary** *if either $p_j = *$ or $p'_j = *$ for all $1 \leq j \leq l$.*

A set of patterns of the same period are said to be *complementary* if every pair of patterns in the set are complementary.

Definition 2.11 *Given a set Π of complementary patterns of the same period l, the* **minimum common superpattern (MCSP)** *of Π is the pattern P of period l, which satisfies the following two conditions.*

- *Each pattern in Π is a subpattern of P.*
- *There does not exist a subpattern P' of P ($P' \neq P$) such that each pattern in Π is also a subpattern of P'.*

It follows from the definition that the information of the MCSP of a set, Π, of complementary patterns is the summation of the information of each pattern in Π.

For a given event segment $D' = d_1, d_2, \ldots, d_l$ and a set, Π, of complementary patterns, the information loss of D' with respect to the MCSP of Π satisfies the following equality

$$L_{D'}(MCSP(\Pi)) = \sum_{P\in\Pi} L_{D'}(P)$$

where $L_D(P)$ is the information loss of D' with respect to P. The rationale is that if D' is not in compliance with a pattern P in Π on position j, then the jth position must be instantiated and D' must not be in compliance with the MCSP of Π on position j either. In general, for any event sequence D, the overall information loss of D with respect to the MCSP of a set of complementary patterns Π is equal to the summation of the information loss of D with respect to each pattern in Π.

Definition 2.12 *A pattern $P = (p_1, p_2, \ldots, p_l)$ is a* **singular pattern** *if only one position is instantiated.*

Intuitively, a singular pattern of period l consists one regular event and $l - 1$ eternal events. The event may occur at one of the l positions. For example, $(a, *, *)$, $(*, a, *)$, and $(*, *, a)$ are all singular patterns of period 3 involving event a.

Proposition 2.1 (Triangle Inequality) *Given an event sequence D and two complementary patterns P and P' of the same period, let Q be the minimum common super pattern of P and P'. Then the generalized information gain of D with respect to Q is at most the summation of that of P and P'.*

Proposition 2.1 can be easily generalized to a set of complementary patterns, which is stated as follows.

Proposition 2.2 *Given an event sequence D and a set of complementary patterns Π, let Q be the minimum common super pattern of Π, then the generalized information gain of D with respect to Q is at most the summation of that of each pattern in Π.*

3 General Approach

In this section, we outline the general strategy we use to mine patterns that meet certain generalized information gain threshold g. There exist three challenges for mining patterns with information gain: (1) The number of different patterns is

$$\sum_{0<l\leq L_{max}} \left(|\Im|^{l} - 1\right) = O(|\Im|^{L_{max}})$$

where $|\Im|$ and L_{max} are the overall number of distinct events and the maximum period length, respectively. Since L_{max} can be quite large, e.g., in the thousands, it is infeasible to verify each pattern against the data directly. Some pruning mechanism has to be developed to circumscribe the search space. (2) By definition, the generalized information gain measure does not have the property of downward closure as the traditional *support* measure does. This prevents us from borrowing existing algorithms developed for association rule problems to mine the qualified patterns. (3) The subsequence concept introduced in this paper poses a difficult challenge to determine when a subsequence should start and end. If a pattern misses some "matches", it is hard to tell whether this signals the end of a subsequence or this merely means some noise within a subsequence.

Fortunately, the *triangle inequality* holds for the generalized information gain. In other word, for a set of complementary patterns Π, the generalized information gain of the minimum common superpattern (MCSP) of Π is always less than or equal to the sum of that of each individual pattern in Π over the same subsequence of events. Inspired by this observation, we can first collect the generalized information gain of all singular patterns, and then generate candidate patterns by combining these singular patterns.

In the first phase, the valid singular patterns are discovered. The second phase generates the candidates of valid complex pattern based on the candidates of valid singular patterns via triangle inequality. Finally, InfoMiner+ verifies all candidates, and finds the corresponding subsequence for each valid pattern so as to maximize its generalized information gain. The first phase can be further divided into two steps: (1) identify the likely periods for each event and (2) find the valid singular pattern for each likely period. In the first step, the negative impact of gap penalties towards the overall GIG is taken into account to prune out disqualified periods of each event; while in the second step all possible format of singular patterns are considered and evaluated. For each likely period l of event a, there are l possible singular patterns, e.g., for $l = 3$, the three singular patterns are $(a, *, *)$, $(*, a, *)$, and $(*, *, a)$. After valid singular patterns are identified, the candidate complex patterns are generated according to the triangle inequality, and then verified. At the same time, for each significant pattern, we will find its associated subsequences. Due to space limitations, we omit the detailed description of the InfoMiner+ algorithm.

4 Conclusion

In this paper, we propose a new mining problem of partial periodic pattern with random replacement. To qualify significant patterns in a sequence, we introduce a new measurement: generalized information gain. This new metric can seamlessly accommodate the different frequency of event occurrences in an event sequence and the gap penalties and provides us with solid theoretical foundations. The triangle inequality preserved by the generalized information gain enables us to devise a linear algorithm to mine the significant pattern in any subsequence combinations.

References

[1] R. Blahut. *Principles and Practice of Information Theory*, Addison-Wesley Publishing Company, 1987.

[2] R. Durbin, S. Eddy, A. Krough, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, 1998.

[3] J. Han, G. Dong, and Y. Yin. Efficient mining partial periodic patterns in time series database. *Proc. Int. Conf. on Data Engineering*, 106-115, 1999.

[4] E. Keogh, S. Lonardi, and W. Chiu. Finding surprising patterns in a time series database in linear time and space. *Proc. ACM Knowledge Discovery and Data Mining*, pp 550-556, 2002.

[5] J. Yang, W. Wang, and P. Yu. Mining asynchronous periodic patterns in time series data. *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 275-279, 2000.

[6] J. Yang, W. Wang, and P. Yu. InfoMiner: mining surprising periodic patterns. *Proc. ACM Knowledge Discovery and Data Mining*, 395-400, 2001.