Revision of "Frequent pattern discovery with tri-partition alphabets"

Fan Min, Zhi-Heng Zhang, Wen-Jie Zhai, Rong-Ping Shen

School of Computer Science, Southwest Petroleum University, Chengdu 610500, China

1. Introduction

We are grateful to the reviewer for the careful review and constructive suggestions with regard to our manuscript. The comments were valuable and very helpful for revising and improving our paper, as well as helping us to refine the direction of our research. We have carefully addressed, as outlined in the following sections, all of the reviewer's suggestions. Particularly, the abstract and the introduction parts are rewritten, and the experimentation is significantly enriched.

2. Response to reviewer #1

The paper introduces a more general and flexible type called tri-pattern by partitioning the alphabet into the strong/medium/weak parts. It indicates an interesting direction of three-way decision. The proposed TPM may be served as a general method for pattern discovery. The connection of TPM and existing methods and its properties are well described.

Response: Thank you for the kind words.

There are a few minor comments:

1. Parameter setting for N, M, etc. should be explained in more detail.

Response: We have revised "Data" subsections to "Data and parameter settings."

2. Instead of talking too much about the experimental results, discussion of further works may be elaborated.

Response: Thank you for the good suggestion. We have added some discussions of further works.

3. English in the experimental part needs improvement.

Response: We have polished the English carefully.

3. Response to reviewer #2

The algorithm the authors propose is powerful.

Response: Thank you for the kind words and so many detailed suggestions.

To enhance a broad readability of the paper, I strongly recommend the authors to add, here and there, some lines of text to help the reader following the sequence of reasoning throughout the paper, especially when considering that examples of different domains are given.

Response: We have added some text especially in the experimental part.

Moreover, the real focus of the paper is dealt with in detail starting from page 17. The abstract should be re-organised and better focussed. The very first sentence is quite circular: "The concept of patterns is the basis of ..pattern discovery" I suggest "The concept of patterns is the basis of sequence analysis".

Response: Suggestion accepted.

It follows with "Inspired by the protein tri-partition " here the reader must receive sufficient information to understand what is the paper about, and which is the purpose. I suggest you something like this: "Inspired by the protein tri-partition and three-way decision, we propose here an algorithm . . . for . . . by dividing the alphabet into strong/medium/weak parts . . . ."

Response: We have accepted this suggestion and revised the sentence as follows. "Inspired by the protein tri-partition and three-way decision, we propose here a frequent pattern discovery algorithm by dividing the alphabet into strong/medium/weak parts."

Definition 3 (page 4 of the manuscript) contains a formal ERROR: "the union of Λ and Ω" should be replaced by "Λ intersects with Ω".

Response: Error corrected.

4.2.1. Data: after equation (13), in line with the idea to help the reader better following the overall reasoning, I suggest the authors to add, for example "To model the fluctuation of oil production, we propose the code table as shown in Table 6".

Response: Suggestion accepted.

4.3. Faked Chinese text: (page 18) the authors give examples of Chinese characters that are part of the Λ set. Here, for a wider audience of readers, it would be useful to give an English translation (whenever possible) of the meanings expressed by the given characters. It follows with the explanation of characters of the weak - Ω - set: English characters are commonly (and historically) defined ad Latin characters. Thus, please change "English" with "Latin".

Response: We have added the translation.

4.3.3. Results: (page 21) The reason is the same as Figure 5(d). Please rephrase it more clearly by adding here a few words to highlight the reason, since a figure is not a reason, it may only show results, trends.

Response: We have revised the sentence as "The reason is that most middle elements in Λ are infrequent."

It follows with bullet points summarising your observations. Point 2: Naturally the accuracy of Type II patterns are 100%. Is it correct? Linguistically, it should be "is 100%". This apart, I cannot see how type II accuracy can be 100%. I suppose it should be "Type I".

Response: By definition, Type II (plain) patterns achieves 100% accuracy when no noise is introduced. However, this cannot be observed from Table 8. Hence we removed this point.

5. Conclusions: The very ending (In the future we will collaborate) of the conclusion should be rephrased. To give but an example: "A prospective application of our TPM algorithm will involve domain-specific experts for some real applications of tri-pattern discovery".

Response: We have extended the conclusion part to include this issue in the first point.

4. Response to reviewer #3

The idea of tri-pattern is interesting.

Response: Thank you for the kind words.

However, there are several problems in this paper in terms of motivations, contributions and clarity. Most of the results come without motivations and comments.

Response: We have tried our best to amend them in this version.

Why we need tri-pattern?

Response: Tri-pattern is a natural expression in many applications. We have discussed three of them in the experimental part.

How the work is inspired by the theory of three-way decisions? What is the relationship between the tri-pattern and the three-way decisions?

Response: 3WD is a methodology rather than a concrete technology. As depicted in Figure 1, the tri-partition and tri-action is inspired by 3WD.

The semantic partition in the alphabet system is universal. Why did the authors choose three-partition rather than five-partition? What is the practical significance of the three-partition? The five-partition is also very common from a qualitative point of view.

Response: This is probably the most frequently asked question about 3WD applications. As you know, any research has its limitations. We choose tri-partition because it is natural for our applications.

Even though this paper includes a large number of references (about 52 items), some of them can be removed since they are not quite related to the topic of this paper.

Response: We have removed them. We also added some in the "Related work" section.

The author had provided a review of other's work in the first paragraph. However, some of references are not the typical work about frequent pattern discovery. Besides, lots of references in the first section are not related to the present work. The introduction should review the history or the existing work on three-way decisions. But, I cannot find the related work about three-way decisions in this paper. Thus, I do not think the work is related to the special issue. I think it is better to add a section like "Related work" to organize more logically the different existing work so that a reader can see clearly the contribution of your paper. So, it is not strange there is lack of the comparative results with the recent work.

Response: It is a good suggestion. We have added this section.

The basic idea of the work is to divide the alphabet into three parts corresponding to strong, medium, and weak characters. But, what are strong characters, medium characters and weak characters? And, different users divide the alphabet differently because they have different criteria. How to deal handle the question? Definition 3 seems cannot support the latter work. I wondered the authors how to get the three partitions in experiments.

Response: You are right. The tri-partition criteria is critical to the approach. For some applications the tri-partition is quite natural, while for others they should be specified by the domain expert. We have added some text to explain our settings for the experiments.

In Definition 4, I think tri-wildcard just is another name of weak-wildcard. There is no much originality.

Response: Their difference is the main contribution of this work. With weak-wildcard, the alphabet is partitioned in two. Type IV and V patterns are based on this existing concept. While with tri-wildcard, the alphabet is partitioned in three. Type I patterns are based on the new concept.

In subsection 2.2, Proposition 1 comes from Ref.[27]. But, Eq. (2) is not in the reference. So, what is means?

Response: Eq. (2) comes from Eq. (4) of the reference. We use different notations to adapt to this new paper.

What is the difference between Definition 7 in Ref. [38] and Definition 5 in this work?

Response: These two types of patterns share the same form. The difference is how the pattern is formed and matched. First, according to our new definition (Definition 5), p_i $(1 \leq i \leq m)$ cannot be a weak character. The existing definition does not have such requirement. Second, the tri-wildcard gap (N, M) in the new definition matches a sequence of medium or weak characters. In contrast, a weak-wildcard gap (N, M) in the existing definition only matches a sequence of weak characters.

Results in Table 5 show that the average values of Type I is not best, please analyze the reasons.

Response: We have added the analysis. Unlike other measures such as predication accuracy, pattern popularity is not as strong in assessing its effectiveness.

The authors should also give the running time of each mode on more data sets.

Response: We have added the running time analysis.

5. Conclusions

We have addressed all of the issues raised by the editor and two reviewers. We hope that the revised manuscript meets the standards for acceptance set by Information Sciences.