Postprint of Differential Sequential Pattern Mining Algorithm Based on Standard Permutation Test

Authors: Wu Jun, Ouyang Aijia, Zhang Lin

Date: 2020-09-28T00:00:00+00:00

Abstract

To eliminate false positive differential sequence patterns from the results returned by differential sequence pattern mining algorithms, we propose an algorithm SP-DSP based on standard permutation hypothesis testing. The algorithm first employs the GSP algorithm to mine frequent sequence patterns, then generates a candidate set of differential sequence patterns based on a Growth rate threshold, and utilizes standard permutation testing to calculate the p-value for each pattern in the candidate set, and finally applies multiple hypothesis testing correction to filter out false positive differential sequence patterns. Experimental results demonstrate that the SP-DSP algorithm can remove a certain number of false positive patterns while preserving true differential sequence patterns as much as possible, thereby improving the accuracy of subsequent classification tasks.

Full Text

Preamble

Vol. 38 No. 3
Application Research of Computers
Accepted Paper

Mining Discriminative Sequential Patterns Based on Standard Permutation Testing

Wu Jun, Ouyang Aijia, Zhang Lin
(School of Information Engineering, Zunyi Normal University, Zunyi, Guizhou 563000, China)

Abstract: To eliminate false positive discriminative sequential patterns from the results returned by discriminative sequential pattern mining algorithms,

this paper proposes a standard permutation hypothesis testing-based algorithm called SP-DSP. The algorithm first employs the GSP algorithm to mine frequent sequential patterns, then generates a candidate set of discriminative sequential patterns based on a Growth rate threshold. It subsequently uses standard permutation testing to calculate the p-value for each pattern in the candidate set, and finally applies multiple hypothesis testing measures to filter out false positive discriminative sequential patterns. Experimental results demonstrate that the SP-DSP algorithm can remove a certain number of false positive patterns while preserving as many true discriminative sequential patterns as possible, thereby improving the accuracy of downstream classification tasks.

Keywords: discriminative sequential pattern mining; pattern assessment; multiple hypothesis testing; standard permutation testing

Mining Discriminative Sequential Patterns Based on Standard Permutation Testing

Wu Jun, Ouyang Aijia, Zhang Lin
(School of Information Engineering, Zunyi Normal University, Zunyi Guizhou 563000, China)

Abstract: To filter out false positive patterns returned from discriminative sequential pattern mining methods, this paper proposes a standard permutation-based method called SP-DSP. This method first mines frequent sequential patterns using the GSP algorithm, then eliminates patterns whose Growth rate is less than the threshold. Finally, the standard permutation method is used to compute the p-values of tested patterns. As a result, the number of false positive patterns can be controlled under multiple hypothesis testing measures. Experiments show that the SP-DSP algorithm can alleviate a large number of false positive patterns and retain as many true patterns as possible, which improves the accuracy of downstream classification tasks.

Key words: discriminative sequential pattern mining; pattern assessment; multiple hypothesis testing; standard permutation testing

0 Introduction

Sequential data refers to data where elements have a certain sequential relationship. Common examples include web browsing sequences, protein sequences, and human language. In sequential data with class labels, certain sequential patterns exhibit significantly different frequencies across different classes. Such patterns are called discriminative sequential patterns [1], which hold considerable value in many applications such as medicine [2–5].

To date, several effective discriminative sequential pattern mining algorithms have been proposed [6–8]. These methods primarily focus on how to quickly and efficiently mine discriminative sequential patterns, but pay little attention to the validity of the discovered patterns. Consequently, the results returned by these algorithms contain a certain number of false positive discriminative

sequential patterns. False positive discriminative sequential patterns refer to patterns that appear randomly in the dataset and do not reflect the true characteristics of the population. Employing such patterns in subsequent research may lead to erroneous results. Statistical significance testing can be used to evaluate the quality of mining results and thereby filter out false positive discriminative sequential patterns.

In statistical significance testing, the significance of a result is measured by its p-value. A smaller p-value indicates stronger statistical significance. When only one discriminative sequential pattern is being tested, if its p-value is less than a threshold α, the pattern is considered statistically significant at significance level α. In many practical situations, multiple discriminative sequential patterns need to be tested simultaneously, a scenario known as multiple hypothesis testing. FWER (family-wise error rate) [9] and FDR (false discovery rate) [10] are two commonly used measures for controlling the number of false positive results in multiple hypothesis testing.

In recent years, using statistical significance testing methods to evaluate data mining results has been extensively studied. For discriminative pattern mining in non-sequential data, Webb proposed two techniques to test pattern validity: the holdout method and the direct computation method [11]. Liu et al. summarized several multiple hypothesis testing algorithms in association rule mining and categorized them into three classes [12], among which permutation testing-based methods are the most effective. Subsequently, several improved permutation testing algorithms have been proposed and applied to discriminative pattern mining [13, 14]. Komiyama et al. proposed two methods, LAMP-EP and QT-LAMP-EP, to control FWER and FDR metrics respectively [15]. These methods have achieved excellent results in discriminative pattern mining for non-sequential data.

Recently, to validate the effectiveness of multiple hypothesis testing for discriminative sequential pattern mining tasks, He et al. designed a direct computation-based method called DSPM-MTC to control the number of false positive patterns [1]. DSPM-MTC directly calculates the p-value for each discriminative sequential pattern based on the property that support follows a hypergeometric distribution. Given that permutation testing methods are more effective than direct computation methods in non-sequential data tasks [12], this paper proposes a discriminative sequential pattern mining algorithm based on standard permutation testing, namely the SP-DSP algorithm. The algorithm first uses the GSP algorithm to mine candidate discriminative sequential patterns [16], then performs standard permutation testing on the original data to obtain the corresponding permutation null distribution, and finally calculates the p-values of candidate discriminative sequential patterns from this null distribution. It employs FWER and FDR measures to control the number of false positive discriminative sequential patterns in the results under statistical significance level α. The main contributions of this paper are as follows:

a) We propose a discriminative sequential pattern mining algorithm SP-DSP

based on standard permutation testing, which can control the number of false positive discriminative sequential patterns in mining results under statistical significance level α.

b) Experimental results on real datasets demonstrate that the SP-DSP algorithm can filter out a certain number of false positive patterns and retain more true discriminative sequential patterns than the DSPM-MTC algorithm.

c) We demonstrate that applying multiple hypothesis testing can improve the credibility of discriminative sequential pattern mining algorithm results.

1.1 Frequent Sequential Patterns

A sequence $s = \langle a_1, a_2, \dots, a_l \rangle$ is an ordered linear list composed of elements from an alphabet $I = \{i_1, i_2, \dots, i_{|I|}\}$, where $a_j \in I$. For sequences $s_1 = \langle a_1, a_2, \dots, a_l \rangle$ and $s_2 = \langle a'_1, a'_2, \dots, a'_m \rangle$, if every element $a'_j \in s_2$ also appears in s_1 and conforms to the element order of s_1, then s_2 is called a subsequence of s_1, denoted as $s_2 \subseteq s_1$. For example, given a sequence $s = \langle i_1, i_3, i_5, i_6, i_8 \rangle$, both $\langle i_1, i_3 \rangle$ and $\langle i_5, i_8 \rangle$ are subsequences of s, while $\langle i_5, i_3 \rangle$ is not a subsequence of s because it does not satisfy the element order of s.

Given a sequential dataset $D = \{t_1, t_2, \dots, t_{|D|}\}$, the support of sequence s in D is defined as the total number of sequences in D that contain s, i.e., $\sup(s, D) = |\{t | t \in D \wedge s \subseteq t\}|$. If the support of a sequence s exceeds a user-defined threshold min_sup, the sequence is called a frequent sequential pattern.

1.2 Discriminative Sequential Pattern Mining

For sequential data with class labels, some sequences exhibit significant frequency differences across different class labels. Such sequences are called discriminative sequential patterns, where the differences across classes can be measured by various discriminative measures [17]. For ease of discussion, we adopt sequential data containing only two class labels, denoted as D^+ and D^- respectively. The proposed method can be easily extended to sequential data with multiple class labels.

Non-statistical discriminative sequential pattern mining algorithms can generally be divided into two steps. First, a frequent sequential pattern mining algorithm is used to mine a certain number of candidate discriminative sequential patterns. Then, the discriminative measure values of these patterns are calculated, and those meeting the given threshold constraints are identified as discriminative sequential patterns.

1.3 Statistical Significance Testing

Statistical significance testing involves two hypotheses: the null hypothesis and the alternative hypothesis. The null hypothesis for discriminative sequential

pattern mining tasks is that discriminative sequential patterns have the same distribution in D^+ and D^- datasets. In this task, the statistical significance of each discriminative sequential pattern is measured by its p-value. The p-value is defined as the probability of obtaining a discriminative sequential pattern that is as extreme as or more extreme than pattern s, assuming that s has the same distribution in D^+ and D^-. The smaller the p-value of a discriminative sequential pattern, the less likely it is to have the same distribution across different classes. When testing a single discriminative sequential pattern independently, if its p-value is less than a threshold α, the pattern is called a statistically significant discriminative sequential pattern at significance level α.

Discriminative sequential pattern mining algorithms typically return a large number of patterns. Using independent testing methods would lead to an increase in false positive results, making this scenario more suitable for multiple hypothesis testing. In multiple hypothesis testing, FWER and FDR are two commonly used statistical measures. FWER is defined as the probability of discovering a false positive discriminative sequential pattern, while FDR is defined as the expected proportion of false positive discriminative sequential patterns. FWER can be controlled using Bonferroni correction [9], and FDR can be controlled using the BH method [10].

2.1 Discriminative Measure

The SP-DSP algorithm adopts Growth rate as the discriminative measure for sequential patterns. Given a frequent sequential pattern s, its Growth rate is calculated as follows: If the Growth rate of a pattern, denoted as $\mathrm{Grow}(s, D)$, is greater than or equal to a user-defined discriminative threshold β, the pattern is called a candidate discriminative sequential pattern.

2.2 Permutation Testing

The standard permutation testing used in SP-DSP consists of five steps:

a) Establish a null hypothesis based on the specific task, select a statistical measure that has different values under the null and alternative hypotheses, and mine candidate discriminative sequential patterns R from the D^+ dataset. The null hypothesis of the SP-DSP algorithm is that discriminative sequential patterns have different distributions in D^+ and D^-, and the selected statistical measure is Growth rate.

b) Randomly exchange sequence data between D^+ and D^- to obtain permuted sequence datasets: D'^+ and D'^-. The permutation process is illustrated in Figure 1 [Figure 1: see original paper]. Suppose D contains 8 sequence data items $\{t_1, t_2, \ldots, t_8\}$, where the first 5 belong to the D^+ dataset and the last 3 belong to the D^- dataset. Randomly generate a permutation sequence: 7, 5, 1, 4, 2, 3, 8, 6. Based on this sequence, assign the label of t_1 to t_7, the label of t_2 to t_5, and so on, to obtain the permuted

sequence dataset.

c) Mine discriminative sequential patterns from the D'^{+} sequence dataset and place the corresponding statistical measure values into set G.

d) After repeating steps 2 and 3 several times, construct the null distribution of this permutation test using the statistical measure values in set G. Typically, 1000 permutations are performed.

e) Place the statistical measure values of candidate discriminative sequential patterns in D^{+} into the above null distribution to calculate the permutation test p-value. The calculation formula is where g_j refers to the statistical measure value of discriminative sequential pattern s on the original dataset.

2.3 Multiple Hypothesis Testing

After obtaining the p-values of candidate discriminative sequential patterns through permutation testing, the SP-DSP algorithm uses Bonferroni correction and the BH method to control the FWER and FDR of the mining results R under statistical significance level α. The FWER calculation formula is When calculating FDR, the p-values of discriminative sequential patterns in R must first be sorted in ascending order to obtain $R' = \{s'_1, s'_2, \ldots, s'_{|R|}\}$, after which we can calculate:

2.4 SP-DSP Algorithm

The pseudocode of the SP-DSP algorithm is shown in Algorithm 1, with detailed explanations as follows:

a) Use the gsp(D^{+}, min_sup) algorithm to mine frequent sequential patterns in the D^{+} sequence dataset with support no less than min_sup, and place them into set Freq (line a). Calculate the Growth rate value for each frequent sequential pattern in set Freq, and place patterns exceeding threshold β into set R. The patterns in R are candidate discriminative sequential patterns (lines b–e).

b) For each permutation j, first use the permutate(D) method to permute class labels and obtain permuted datasets D'^{+} and D'^{-}. Then, use the gsp(D'^{+}, min_sup) algorithm to mine frequent sequential patterns in D'^{+} and place them into set Freq_per$_j$. Next, use the com_sta(Freq_per$_j$, β) method to calculate the Growth rate value for each frequent sequential pattern in Freq_per$_j$, and place values exceeding β into set G_j. Finally, merge the Growth rate values in G_j into set G (lines f–k). All Growth rate values in G constitute the null distribution of this permutation test.

c) Place the discriminative measure value of each candidate discriminative sequential pattern in set R into the null distribution to calculate its p-value (lines l–n). Then, filter out non-redundant candidate discriminative

sequential patterns to obtain set R' based on each pattern' s p-value (line o). Finally, use Bonferroni correction to control the FWER of R' under statistical significance level α, and save statistically significant discriminative sequential patterns into set R'_{FWER}. Similarly, use the BH method to control the FDR of R' under statistical significance level α, and save statistically significant discriminative sequential patterns into set R'_{FDR} (lines p-q).

Algorithm 2 describes the detailed steps of the redundancy removal method redundancy_filter(R): For each candidate discriminative sequential pattern r, first find its corresponding subpattern set Sub. Then find the minimum p-value min_p among them. If r' s p-value is less than min_p, place r into the non-redundant candidate discriminative sequential pattern set R'. Finally, return R' for subsequent evaluation.

2.5 Experimental Data

Four sequence datasets of different sizes were selected for the experiments: Linux1_5 [18], Question [19], WebKB [20], and Reuters [21]. Among them, Linux1_5, WebKB, and Reuters are multi-class datasets; in the experiments, only the two classes with the largest number of sequences were retained from these three datasets. Detailed dataset information is shown in Table 1 , where l_{min}, l_{max}, and l_{avg} represent the minimum, maximum, and average sequence lengths, respectively.

2.6 Experimental Results

To evaluate the performance of the SP-DSP algorithm, extensive comparative experiments were conducted on real datasets. The comparison algorithms are the IMP algorithm [6], CGM algorithm [8], and DSPM-MTC algorithm [1]. DSPM-MTC is a discriminative sequential pattern mining algorithm based on multiple hypothesis testing, while IMP and CGM are discriminative sequential pattern mining algorithms based on discriminative measures. Both IMP and CGM use the redundancy removal method employed in [1]. All experiments were run on a computer with a 2.40 GHz CPU and 12 GB memory.

The experiments first compared the number of discriminative sequential patterns returned by SP-DSP$_{\mathrm{FDR}}$, DSPM-MTC$_{\mathrm{FDR}}$, IMP, and CGM on different datasets under the same min_sup, α, and β parameters. The results are shown in Figure 2 [Figure 2: see original paper], which clearly demonstrates that SP-DSP$_{\mathrm{FDR}}$ and DSPM-MTC$_{\mathrm{FDR}}$ return fewer patterns than IMP and CGM. This is because discriminative measure constraints only focus on the patterns themselves, while multiple hypothesis testing evaluates the entire set of returned results, making algorithms based on multiple hypothesis testing impose stricter quality constraints on the results.

Subsequently, the experiments compared the number of discriminative sequen-

tial patterns returned by SP-DSP and DSPM-MTC under FWER and FDR constraints with the same parameters. The results are shown in Table 2 . The results indicate that under the same FDR or FWER constraint, SP-DSP returns more patterns than DSPM-MTC, demonstrating that permutation testing-based methods can report more results than direct computation-based methods. Additionally, the same method reports fewer discriminative sequential patterns under FWER constraint than under FDR constraint, proving that FWER imposes stricter constraints than FDR.

These experimental results show that compared with non-multiple hypothesis testing methods IMP and CGM, SP-DSP can filter out a large number of patterns; compared with the multiple hypothesis testing method DSPM-MTC, SP-DSP can retain more discriminative sequential patterns.

Since real datasets lack ground truth information about discriminative sequential patterns, the above results cannot directly demonstrate that SP-DSP finds more accurate patterns than other methods. To prove the accuracy of the mining algorithm, subsequent experiments used the mined patterns as features for classification prediction tasks [22]. Classification tasks can validate pattern accuracy because true discriminative sequential patterns reflect the distributional differences between different class datasets, which correspond to class labels. Specifically, based on the number of mined discriminative sequential patterns, each sequence in the dataset was constructed into a vector of the same size as a feature representation. If a sequence contains a particular pattern, the value for that feature is 1; otherwise, it is 0.

Considering the impact of different classification methods, three classification methods with different mechanisms were used: Naive Bayes (NB), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). To avoid randomness, each classification method used five-fold cross-validation, and the average of ten prediction results was taken as the final classification accuracy. The detailed experimental results are shown in Tables 3 –5.

From the experimental results of the three classification methods, we can observe: First, the classification accuracy using features constructed from SP-DSP_{FDR} and DSPM-MTC_{FDR} results is significantly higher than that using features from IMG and GCM results, demonstrating that multiple hypothesis testing indeed filters out many false positive discriminative sequential patterns. Taking the Question dataset as an example, the IMG and GCM mining results contain the pattern $\langle \text{where}, \text{the} \rangle$, while SP-$\text{DSP}_{\text{FDR}}$ and DSPM-MTC_{FDR} results only contain $\langle \text{where} \rangle$. Many sequences in the Question dataset contain the definite article "the," which has no substantive meaning, so using it as a feature is likely to cause misclassification.

Second, the accuracy of SP-DSP_{FDR} results is higher than that of DSPM-MTC_{FDR} results, indicating that the additional discriminative sequential patterns retained by permutation testing are likely true patterns. Taking the Question dataset as an example, SP-DSP_{FDR} results contain patterns $\langle \text{what} \rangle$ and

$\langle \text{what}, \text{in} \rangle$, while DSPM-MTC$_{\text{FDR}}$ results only contain $\langle \text{what} \rangle$. Observing the final misclassification results reveals that DSPM-MTC$_{\text{FDR}}$ misclassifies 6 sequences containing the $\langle \text{what}, \text{in} \rangle$ pattern from the positive class to the negative class, while SP-DSP$_{\text{FDR}}$ can classify all of them correctly, suggesting that $\langle \text{what}, \text{in} \rangle$ is likely a true discriminative pattern.

Furthermore, the low accuracy of GCM on the WebKB and Reuters datasets demonstrates the severe misleading effect of false positive patterns on downstream tasks. Meanwhile, IMG and GCM algorithms are more sensitive to different classification methods because they generate more interfering features.

3 Conclusion

To improve the accuracy of discriminative sequential pattern mining tasks, we propose the SP-DSP algorithm based on standard permutation testing. Experimental results on real datasets prove that applying multiple hypothesis testing can enhance the credibility of discriminative sequential pattern mining algorithms. Compared with the DSPM-MTC method, the SP-DSP algorithm can retain as many true discriminative sequential patterns as possible. Due to the randomness of standard permutation testing, the number of statistically significant discriminative sequential patterns returned by the SP-DSP algorithm may fluctuate. This experiment uses the average of ten algorithm runs as the final result; future work will investigate the selection problem of borderline patterns.

References

[1] He Zengyou, Zhang Simeng, Wu Jun. Significance-based discriminative sequential pattern mining [J]. Expert Systems with Applications, 2019, 122(1): 54-64.

[2] Ghosh S, Li Jinyan, Cao Longbin, et al. Septic shock prediction for ICU patients via coupled HMM walking on sequential contrast patterns [J]. Journal of Biomedical Informatics, 2017, 66(1): 19-31.

[3] 高增, 郭均鹏. 考虑价格的跨种类模糊序列模式挖掘算法 [J]. 计算机应用研究, 2018, 35(1): 39-42, 47. (Gao Zeng, Guo Junpeng. Cross-category fuzzy sequential pattern mining algorithm with consideration of price [J]. Application Research of Computers, 2018, 35(1): 39-42, 47.)

[4] 张光兰, 杨秋辉, 程雪梅, 等. 序列模式挖掘在通信网络告警预测中的应用 [J]. 计算机科学, 2018, 45(S2): 535-538, 563. (Zhang Guanglan, Yang Qiuhui, Cheng Xuemei, et al. Application of sequence pattern mining in communication network alarm prediction [J]. Computer Science, 2018, 45(S2): 535-538, 563.)

[5] Liu Lu, Duan Lei, Yang Hao, et al. Mining distinguishing customer focus sets for online shopping decision support [C]//Proc of the 12th International Conference on Advanced Data Mining and Applications. Switzerland: Springer press, 2016: 50-64.

[6] Zheng Zhigang, Wei Wei, Liu Chunming, et al. An effective contrast sequential pattern mining approach to taxpayer behavior analysis [J]. World Wide Web, 2016, 19(4): 633-651.

[7] He Zengyou, Gu Feiyang, Zhao Can, et al. Conditional discriminative pattern mining: concepts and algorithms [J]. Information Sciences, 2017, 375(1): 1-15.

[8] Ji Xiaonan, Bailey J, Dong Guozhu. Mining minimal distinguishing subsequence patterns with gap constraints [J]. Knowledge and Information Systems, 2007, 11(3): 259-286.

[9] Bland J, Altman D. Multiple significance tests: The bonferroni method [J]. British Medical Journal, 1995, 310(6): 170-176.

[10] Benjamini Y, Krieger A, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate [J]. Biometrika, 2006, 93(3): 491-507.

[11] Webb G. Discovering significant patterns [J]. Machine Learning, 2007, 68(1): 1-33.

[12] Liu Guimei, Zhang Haojun, Wong L. Controlling false positives in association rule mining [J]. Proceedings of the VLDB Endowment, 2011, 5(2): 145-156.

[13] Leonardo P, Fabio V. Efficient mining of the most significant patterns with permutation testing [C]//Proc of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. London: ACM press, 2018: 2070-2079.

[14] 吴军, 段琼, 张琳, 等. 磷酸化基序精确置换检验 p-value 的计算方法 [J]. 中国科学: 信息科学, 2017, 47(10): 1334-1348. (Wu Jun, Duan Qiong, Zhang Lin, et al. Computing exact permutation p-values for phosphorylation motifs [J]. Scientia Sinica Informationis, 2017, 47(10): 1334-1348.)

[15] Komiyama J, Ishihata M, Arimura H, et al. Statistical emerging pattern mining with multiple testing correction [C]//Proc of the 23th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax: ACM press, 2017: 897-906.

[16] Srikant R, Agrawal R. Mining sequential patterns: generalizations and performance improvements [C]//Proc of the 6th International Conference on Extending Database Technology. Heidelberg: Springer press, 1996: 1-17.

[17] Liu Xiaoqing, Wu Jun, Gu Feiyang, et al. Discriminative pattern mining and its applications in bioinformatics [J]. Briefings in Bioinformatics, 2015, 16(5): 884-900.

[18] Dua D, Graff C. UCI machine learning repository [EB/OL]. (2007) [2018-09-24]. http://archive.ics.uci.edu/ml.

[19] Kim Y. Convolutional Neural Networks for Sentence Classification [C]//Proc of the 18th Conference on Empirical Methods in Natural Language Processing. Doha: ACL press, 2015: 1745-1751.

[20] Craven M, Dipasquo D, Freitag D, et al. Learning to construct knowledge bases from the world wide web [J]. Artificial Intelligence, 2000, 118(1): 1-62.

[21] Cardoso C. Improving methods for single-label text categorization [D]. Lisboa: Instituto Superior Técnico, 2007.

[22] Fradkin D, Mörchen F. Mining sequential patterns for classification [J]. Knowledge and Information Systems, 2015, 45(3): 731-749.

Note: Figure translations are in progress. See original paper for figures.