Proceedings Paper:

https://doi.org/10.1109/ICNIT.2010.5508469
Sequential Data Mining using Correlation Matrix Memory

Sanil Shanker KP
Department of Computer Science, University of Kerala, India
e-mail: sanilshankerkp@gmail.com

Elizabeth Sherly
Indian Institute of Information Technology and Management- Kerala, India

Abstract: This paper proposes a method for sequential data mining using correlation matrix memory. Here, we use the concept of the Logical Match to mine the indices of the sequential pattern. We demonstrate the uniqueness of the method with both the artificial and the real datum taken from NCBI databank.

Keywords: Correlation matrix memory, Logical Match, Sequential data mining.

I INTRODUCTION

In this paper we put forward a unique method of search for sequential patterns using correlation matrix memory by Logical Match. The correlation matrix memory is a particular type of binary associative neural network (Austin,1994). In Logical Matching strategy, the sequence is arranged so that each element in the pattern coincides with it's corresponding index and then proceed to match logically the indices of the subsequence with those of the text pattern (Sanil et al,2010). A correlation matrix memory can be represented as a m X n binary matrix D, where m and n are input width and output width respectively. For a given sequential pattern the input binary vector Isp operates logical AND with reference matrix (database D) gives Isp'. During the transformation, Op ← Ip . D, we mine the indices of both the pattern and text using logical matching, where Isp' = Osp, the output binary vector of the sequential pattern.

II METHOD

1. Initialize Text and Pattern
2. Initialize input binary vector Ip ← Pattern and I_T ← Text
3. Create the reference database D
4. Select (indices of Pattern) when Op ← Ip . D
5. Select (indices of Text) when O_T ← I_T . D
6. Match (indices of Pattern) with (indices of Text) in the increasing order of indices.

III SIMULATION WITH ARTIFICIAL DATUM

Text => <CTCACTCCTC>
Pattern => <CTC>

Initialize {0001← A, 0010← T, 0100← G, 1000← C}

Shift the Text(Table 1) so that the input binary vector I_{Text} operates logical AND with reference matrix ( Database D) gives I_{Test}^T (Table 2).

Input Text (I_{Text})

Database D

<table>
<thead>
<tr>
<th>C</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>C</th>
<th>T</th>
<th>C</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>AND</td>
<td>0/1</td>
<td>0/1</td>
<td>0/1</td>
<td>0/1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1.

<table>
<thead>
<tr>
<th>w</th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2. Transpose of Input Text I_{Text}^T

978-1-4244-7578-0/$26.00 © 2010 IEEE 470
That is, indices of the Text, <w(7);

z(1,3,4,6,8,10)>.

Shift the Pattern (Table 3) so that the input binary vector

\[ \text{Database D} \]

(Table 4) gives \[ \text{Pattern CTC} \] (Table 4). Input Pattern (\( I_{\text{pattern}} \))

Locating Pattern in the Text:

\[ z: 1 \ 3 \ 4 \ 6 \ 8 \ 10 \]

That is, indices of the pattern, \(<x(2): z(1,3)>\).

Here, Pattern CTC is repeating in the locations (1,2,3);(4,5,6);(8,9,10) of the text.

IV EXPERIMENTAL RESULTS

For testing the proposed method, the program has been written in C++ language under Linux platform. The method was tested against DNA sequences of various sizes taken from NCBI databank (Table 5). The method provides the solution to problem of locating the exact position of the pattern in the text.

<table>
<thead>
<tr>
<th>Location of Pattern in Text</th>
</tr>
</thead>
</table>
| \( (16,17,18);(53,54,55); (59,60,61);(71,72,73); (101,102,103);(111,112,113); (136,137,138);(142,143,144); (153,154,155);(159,160,161); (175,176,177);(363,364,365); (372,373,374);(386,387,388); (441,442,443);(450,451,452); (463,464,465);(533,534,535); (554,555,556);(562,563,564); (663,664,665);(685,686,687); (689,690,691);(697,698,699) \)

<table>
<thead>
<tr>
<th>Kennedy disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 3 \ 0 \ 0 \ 0 \ 1 )</td>
</tr>
<tr>
<td>2 \ 0 \ 1 \ 0 \ 0</td>
</tr>
<tr>
<td>1 \ 0 \ 0 \ 0 \ 1</td>
</tr>
</tbody>
</table>

IV EXPERIMENTAL RESULTS

For testing the proposed method, the program has been written in C++ language under Linux platform. The method was tested against DNA sequences of various sizes taken from NCBI databank (Table 5). The method provides the solution to problem of locating the exact position of the pattern in the text.
We present a new sequential data mining method using correlation matrix memory. Here, we use the concept of Logical Match to locate the pattern in the text. This method can possibly be implement to develop a new approach related to the sequential data mining.

ACKNOWLEDGEMENT

SSKP was funded in part by European Research and Educational Collaboration with Asia

REFERENCES

