Chapter V

Mining Progressive Confident Rules in Sequence Databases

co-authored with

Minghua Zhang, National University of Singapore, Singapore

Real-life objects can be described by its attribute values. For example, a person has attributes such as gender, date of birth, education level, and job, and so forth. While the gender and date of birth of a person do not change, the education level and job may change with time. If we denote the set of attribute values of an object as its state, then the state of an object changes as the attribute values change with time. The states of an object at different time stamps form a state sequence.

In many applications, an object’s state sequence over time is usually more interesting than its current state because the state sequence depicts the object’s behavior characteristics. For example, if we only look at the current stock price, we cannot tell its behavior in the next week. However, if we look at its price history over several months, we could have a better idea about its trend in the future.

The state sequence of an object also captures more information than the current state for classification. For example, it is hard to determine whether a patient has chronic lymphocytic leukemia if we only know that he currently
has anemia. However, if we track the patient’s medical history over a period of time, then the doctor will be able to make a better judgement.

Table 5.1 shows a sample database that records the symptom sequences of patients and whether they have Chronic Lymphocytic Leukemia (CLL). The first column is the patient’s ID. The second column is the patient’s symptom sequence over time. For example, for patient ID “1”, he/she first has night sweat and hypodynamia, followed by fever, then achroacytosis, and finally anemia. We regard the set of symptoms at a time point as the state of a patient at that time. Thus, the first state of patient 1 is \{'night sweat, hypodynamia\}', and his last state is anemia. The last column in Table 5.1 is the doctor’s diagnosis on Chronic Lymphocytic Leukemia (CLL). The first four patients have CLL, while the last four do not.

If we examine the last state of each patient, or their last symptom, we find that three patients with anemia have CLL (IDs 1-3), while the other three patients with anemia do not (IDs 6-8). Thus it is not clear whether a patient with anemia will have CLL. However, if we study the patients’ state sequences, or their sequence of symptoms, we see that most CLL patients (IDs 1-3) have the following symptom sequence which does not occur in non-CLL patients: night sweat \(\rightarrow\) fever \(\rightarrow\) achroacytosis \(\rightarrow\) anemia.

We observe that patient 5 in Table 5.1 does not have CLL although his symptom sequence is the same as the 3 leading symptoms of CLL patients (night sweat \(\rightarrow\) fever \(\rightarrow\) achroacytosis). It is possible that he may have CLL later, and a doctor could advise on preventive measures in advance.

### Table 5.1. Example diagnosis on chronic lymphocytic leukemia

<table>
<thead>
<tr>
<th>Patient</th>
<th>Sequence of Symptoms</th>
<th>CLL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{'Night sweat, hypodynamia} \rightarrow Fever \rightarrow Achroacytosis \rightarrow Anemia</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>Night sweat \rightarrow Fever \rightarrow Achroacytosis \rightarrow Anemia</td>
<td>Y</td>
</tr>
<tr>
<td>3</td>
<td>Night sweat \rightarrow Fever \rightarrow Achroacytosis \rightarrow Anemia</td>
<td>Y</td>
</tr>
<tr>
<td>4</td>
<td>Night sweat \rightarrow Fever \rightarrow Achroacytosis \rightarrow Splenomegalias</td>
<td>Y</td>
</tr>
<tr>
<td>5</td>
<td>Night sweat \rightarrow Fever \rightarrow Achroacytosis</td>
<td>N</td>
</tr>
<tr>
<td>6</td>
<td>Night sweat \rightarrow Fever \rightarrow Anemia</td>
<td>N</td>
</tr>
<tr>
<td>7</td>
<td>Night sweat \rightarrow Splenomegalias \rightarrow Anemia</td>
<td>N</td>
</tr>
<tr>
<td>8</td>
<td>Night sweat \rightarrow Sleepy \rightarrow Anemia</td>
<td>N</td>
</tr>
</tbody>
</table>
Related Content

A Graph-Based Biomedical Literature Clustering Approach Utilizing Term's Global and Local Importance Information
[www.igi-global.com/article/graph-based-biomedical-literature-clustering/1819?camid=4v1a](www.igi-global.com/article/graph-based-biomedical-literature-clustering/1819?camid=4v1a)

SQL-Based Fuzzy Query Mechanism Over Encrypted Database
[www.igi-global.com/article/sql-based-fuzzy-query-mechanism-over-encrypted-database/117159?camid=4v1a](www.igi-global.com/article/sql-based-fuzzy-query-mechanism-over-encrypted-database/117159?camid=4v1a)

Text Mining in Bioinformatics: Research and Application
[www.igi-global.com/chapter/text-mining-bioinformatics/21755?camid=4v1a](www.igi-global.com/chapter/text-mining-bioinformatics/21755?camid=4v1a)
Bayesian Data Mining and Knowledge Discovery
www.igi-global.com/chapter/bayesian-data-mining-knowledge-discovery/7604?camid=4v1a