Chapter 14
A New Similarity Metric for Sequential Data

Pradeep Kumar
Indian Institute of Management, India

Bapi S. Raju
Infosys Technologies Limited, India

P. Radha Krishna
University of Hyderabad, India

ABSTRACT

In many data mining applications, both classification and clustering algorithms require a distance/similarity measure. The central problem in similarity based clustering/classification comprising sequential data is deciding an appropriate similarity metric. The existing metrics like Euclidean, Jaccard, Cosine, and so forth do not exploit the sequential nature of data explicitly. In this chapter, the authors propose a similarity preserving function called Sequence and Set Similarity Measure ($S^3M$) that captures both the order of occurrence of items in sequences and the constituent items of sequences. The authors demonstrate the usefulness of the proposed measure for classification and clustering tasks. Experiments were conducted on benchmark datasets, that is, DARPA'98 and msnbc, for classification task in intrusion detection and clustering task in web mining domains. Results show the usefulness of the proposed measure.

INTRODUCTION

Sequential data may arise from diverse application domains which may have time stamp associated with it or not (Salva & Chakravarthy, 2008). They may be music files, system calls, transaction records, web logs, genomic data and so on. In these data there are hidden relations that should be explored to find interesting information. For example, from web logs one can extract the information regarding the most frequent access path; from genomic data one can extract letter or motif (sequence of letters) frequencies; from music files one can discover harmonies etc. One can extract features from sequential data, represent them as vectors and cluster the data using exist-
ing clustering techniques. Similar to clustering, in classification task also a similarity measure is required to determine the class membership of test data or sequence. The central problem in similarity based classification/clustering problems is to come up with an appropriate similarity metric.

Usually when dealing with sequences, we first convert them into frequency vectors, treating all the events within the sequences as independent of one another. The resulting vectors corresponding to the data are then classified/clustered using existing classification/clustering techniques (Tan et al., 2006; Kumar et al., 2007). Treating sequences in this manner results in a loss of the sequential information embedded in them and leads to inaccurate classification or clustering.

A number of metrics have been proposed for sequences, many of them do not really qualify as metrics, as they do not satisfy one or more of the requirements of being a metric (Mitchell, 1997). Similarity has both a quantitative and a qualitative aspect. Some measures such as cosine similarity, hamming distance consider only the quantitative aspect whereas measures such as Longest Common Subsequence (LCS), feature distance consider only qualitative aspect. In this chapter, we introduce a new similarity measure that considers both sequence (qualitative or ordering aspect) and set similarity (quantitative aspect) among sequences while computing similarity. We tested the performance of our proposed similarity measure on both classification and clustering tasks. Standard algorithms like k-Nearest Neighbor (kNN) classification and Partitioning Around Medoid (PAM) clustering algorithms were used along with the cosine measure as well as the proposed similarity measure for comparative experimental analysis. In addition, in the case of classification task, the proposed measure was also compared with a recently proposed metric called, the Binary Weighted Cosine (BWC) similarity measure (Rawat et al., 2006). The effectiveness of the proposed measure is studied in both intrusion detection (classification task) and in web usage mining (clustering task).

This chapter is organized as follows. In the next section, we discuss various aspects of sequence similarity. In the proposed measure, Longest Common Subsequence is one of the components therefore we provide study of longest common sub-sequence in the following section. A new similarity measure $SM$ in is presented in the next followed section. Last but not the final section we present the results of the new measure for both classification and clustering tasks.

**SEQUENCE SIMILARITY**

A sequence is made of set of items that happen in time, or happen one after another, that is, in position but not necessarily in relation with time. We can say that a sequence is an ordered set of items. A sequence is denoted as follows:

$$S = <a_1, a_2, ..., a_n>$$

where $a_1, a_2, ..., a_n$ are the item sets in sequence $S$. Sequence $S$ contains $n$ elements or ordered item sets. Sequence length is defined as the count of number of item sets contained in the sequence. It is denoted as $|S|$ and here, $|S| = n$. Formally, similarity is a nonnegative real valued function $S$, defined on the Cartesian product $X \times X$ of a set $X$. It is called a metric on $X$ if for every $x, y \in X$, the following properties are satisfied by $S$.

1. Non-negativity: $S(x, y) \geq 0$
2. Symmetry: $S(x, y) = S(y, x)$
3. Normalization: $S(x, y) \leq 1$

A set $X$ along with a metric is called a metric space.

Sequence mining algorithms make use of either distance functions (Duda et al., 2001) or similarity functions (Bergroth et al., 2000) for comparing pairs of sequences. Sequence comparison finds
Related Content

A Parameterized Framework for Clustering Streams
www.igi-global.com/article/parameterized-framework-clustering-streams/1822?camid=4v1a

Data Mining Applications in the Electrical Industry
www.igi-global.com/chapter/data-mining-applications-electrical-industry/73514?camid=4v1a

Algebraic and Graphic Languages for OLAP Manipulations
www.igi-global.com/article/algebraic-graphic-languages-olap-manipulations/1798?camid=4v1a

Data Mining for Human Resource Information Systems
www.igi-global.com/chapter/data-mining-human-resource-information/7609?camid=4v1a