# A New Similarity Metric for Sequential Data

*Pradeep Kumar, Indian Institute of Management Lucknow, India*

*P. Radha Krishna, Infosys Technologies Limited, Hyderabad, India*

*Bapi S. Raju, 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 paper, 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.*

*Keywords:* *Sequence Classification, Sequence Clustering, Sequence Data, Similarity Measures, Similarity Metric*

## 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 existing 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

DOI: 10.4018/jdwm.2010100102

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

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 paper, 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 Mediod (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 paper 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 $S^3M$ 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 important and interesting applications in the field of computer science both from the theoretical as well as practical points of view. A wide variety of applications of sequence similarity is seen in various interrelated disciplines such as computer science, molecular biology, speech and pattern recognition etc. Sankoff and Kruskal (1983) present the application of sequence comparison and various methodologies adopted in the literature. In computer science, sequence comparison finds application in various fields such as string matching, classification of imbalance data (Nikulin, 2008) and clustering. Mohamad et al (2006) proposes a similarity retrieval algorithm for time series data.

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

15 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/article/new-similarity-metric-sequential-data/46941

## Related Content

A TOPSIS Data Mining Demonstration and Application to Credit Scoring
Desheng Wu and David L. Olson (2006). *International Journal of Data Warehousing and Mining (pp. 16-26).*
www.irma-international.org/article/topsis-data-mining-demonstration-application/1768/

Data Mining Techniques for Web Personalization: Algorithms and Applications
Gulden Uchyigit (2010). *Dynamic and Advanced Data Mining for Progressing Technological Development: Innovations and Systemic Approaches (pp. 1-17).*
www.irma-international.org/chapter/data-mining-techniques-web-personalization/39634/

Identify Cross-Selling Opportunities via Hybrid Classifier
Dahong Qiu, Ye Wang and Bin Bi (2008). *International Journal of Data Warehousing and Mining (pp. 55-62).*
www.irma-international.org/article/identify-cross-selling-opportunities-via/1807/

Design and Implementation of Active Stream Data Warehouses
Sandro Bimonte, Omar Boussaid, Michel Schneider and Fabien Ruelle (2019). *International Journal of Data Warehousing and Mining (pp. 1-21).*
www.irma-international.org/article/design-and-implementation-of-active-stream-data-warehouses/225804/

White Patch Detection in Brain MRI Image Using Evolutionary Clustering Algorithm
Pradeep Kumar Mallick, Mihir Narayan Mohanty and S. Saravana Kumar (2016). *Research Advances in the Integration of Big Data and Smart Computing (pp. 323-339).*
www.irma-international.org/chapter/white-patch-detection-in-brain-mri-image-using-evolutionary-clustering-algorithm/139410/