JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Novel Sequence Data Similarity Finding Technique for Sequence Data Classification and Clustering

Dr. D. Mabuni
Assistant Professor
Dept. of Computer Science and Technology
Dravidian University, Kuppam, Andhra Pradesh, India

Abstract: Sequence data classification and clustering are important operations in machine learning. A quantitative similarity measure is needed to find the similarity between two given input sequences before applying classification and clustering operations of sequence data records. Many well defined and well-designed functional formulas are already available for sequence data similarity measure computations. In this paper a new functional formula is proposed for sequence data similarity finding. This new sequence data similarity finding measure is based on total number of common subsequences present in the given two input sequences and the satisfaction of longest common subsequence (LCS) property of each subsequence present in the given input data sequences. LCS preserves not only the ordering information of symbols but also correctly maintains subsequence relationship details. Proposed formula is simple to use and compute similarity between input pair of sequence data records. This formula is named as Sequence Data Similarity (SDS).SDS similarity measure computation is completely based on given pair of sequences and all possible common subsequences in both the input pair of sequences.

Index Terms: **Sequence classification, sequence data classification, sequence data clustering,sequence similarity finding measure, sequential data pattern mining, sequential pattern matching, data sequence classification, sequential data analysis, machine learning.**

I. INTRODUCTION

Sequence data classification is the task of predicting a class label for a given data sequence of symbols. Sequence data operations such as sequence data classifications and sequence data clustering are predominantly used in many real-world applications. Sequence similarity finding measurements play an important role in sequence data modeling-based operations such as sequence data classification, sequence dataclustering and sequence datasegmentation. A number of research papers have been presented on sequence data classification and clustering. In general, sequences are transformed into vectors and then existing classification and clustering algorithms such as decision trees, support vector machines (SVM), K-nearest neighbor classifiers, Bayes classifcation and Naïve Bayes are used for sequence data classification.Interpretability properties such as explanation, easy understanding, simplicity, verification, validity, modifiability, applicability and usability are the fundamentally required features of any classifier including sequence data classifiers.One simple way for sequence data classification and clustering is that given set of data sequences are transformed into set of fixed length vectors before sequence data classification or clustering. This vector transformation must be optimally correct in the sense that whenever vectors are re-constructed into the original input data sequences the reconstruction error must be negligible or it must be very close to zero error value. Sequential data analysis,sequential data pattern mining, and sequential pattern matching operations are frequently computed for efficient and effective management of sequence data records. A complex symbolic sequence is defined as an ordered list of vectors.

Cluster quality is directly depending on the similarity finding measure used for sequence data comparisons. Identifying and then applying the correct sequence data similarity measure for categorical sequence data clustering is the biggest challenging task. Main requirement of sequence data processing is that of generating the correct sequential output for a given sequential data input. In the case of parts-of-speech tagging the input is a sequence of words (sentence) and the output is also a sentence with similar meaning. Spatiotemporal sequence data modeling is also needed in addition to the conventional sequence data modeling. Missing values and misalignments are the two main problems associated with Spatiotemporal sequence data modeling. Non-parametric predictors such as decision tree models are very much useful for spatiotemporal sequence prediction but decision trees are not able to handle variable length sequences. One possibility is first complete translating variable length sequences into fixed length sequences and then apply decision tree classifiers or random forests for prediction. Decision tree predictions are made via traversing from the root node of the decision tree to the leaf node of the decision tree.

Basically, sequence data classification is a predictive modeling problem where a set of input data sequences are given and the actual real required task is to find the class label or category of the test sequence data input record. Sequence data classification is difficult when input

data sequences are of variable length. For sequence data classification sequence data learning procedure is required. Analysis of discrete data sequences is an important task in sequence data mining. Various effective methods have been developed and proposed for sequence data classification and clustering operations. Reference or landmark points are potentially useful points in sequence data clustering and classifications. In the case of mining of interesting patterns in the sequences one must consider and must be taken into consideration the values of support and cohesion measures. These interesting patterns in the sequences are used for building sequence data classifiers. A number of sequences data-based classifiers have been developed in the literature.Many sequence data classification techniques already have been proposed for applying in the selected and specific domains.Once the task of discovering all interesting and wanted patterns in each class of sequences then the next step is to identify the actual data classification rules that are needed to build a classifier model. Sometimes there is a need to generate interesting sub-sequences or sub patterns by using already existing algorithms. Also, one must remember that various parameters must be considered and taken into consideration for sequence data classification, sequence data clustering and sequence data patten mining.Sometimes a sequence may carry a specific class label.

The aim of machine learning is to analyze large amounts of data using special computer models and then applying data processing operations efficiently and effectively for getting desired output results. Combination of machine learning models and statistical techniques, hybrid data modeling techniques,can identify data patterns and sequence data patterns present in the input data. Sequential data patterns mining is one of the important and frequently needed operations in the machine learning applications. In general, a sequence data is an ordered list of symbols or elements. Tremendous changes have been brought after the invention of decision trees such as ID3, C4.5, CART and so on. These decision trees, both classification and regression trees, are continuously being used in many real time applications such as retail, medicine, business, finance, banking, and health sciences and so on. So, for all of the existing popular tree-based classifiers which are considered to be of-the-self data classifiers and are useful only for processing of fixed length records or vectors. So, there is a need to extend all these already existing classifiers for handling not only fixed length input records but also originally given actual variable length records of input data sequences.Sequence data classification has an extensive range of applications in real time situations of day to day operations. Support vector method (SVM) is considered to be the most effective method for protein sequence data classification. In many sequence data applications, the basic fundamentally needed operations are sequence data classification of symbolic sequences and time series data. The problem of classifying complex data sequences is still an open problem in sequence data management. Streaming sequence data classification requirements are increasing but streaming data sequence data management is very difficult. Time series sequence data modelling tasks generally include classification, regression, and forecasting. Classification step basically includes model tuning, model training, and model validation. In machine learning different sequence data modelling algorithms are – shape based, statistical modelling, distance based, index based, and tree-based algorithms.

Sequence data classification is one of the important tasks in machine learning. Sequence data classifies play an important role in many day-to-day and real-world applications of many domains. Sequence data classification is learning a data sequence classifier model that maps a sequence to a class label. Out of all the many sequence data classification methods pattern-based sequence data classification method is popular and predominantly used in rea time applications. There is an urgent need to find a very good and highly general and standard frame-worksfor efficient and effective classifying and clustering of any type of data sequences. In general, a sequence is considered as an ordered list of elements or objects. There exist many sequential data analysis tasks in the real time situations. Ordering of different elements in the sequence is important and necessary but it is very difficult to maintain the ordering of elements. Sequential data can be clustered, classified, or subjected to find various desired data patterns. Classification is a method of assigning a class label to the given new or unknown data object, test record. Sequence data classification is defined as assigning the correct class label to the new sequence by the selected efficient classifier constructed from a given training dataset of sequences of given input data records.

Sequence data classification based on interesting patterns is also increasing in use. Basically, sequence data classification techniques are divided into three types:

1) Feature based sequence classification techniques
2) Distance based sequence classification techniques and
3) Model based sequence classification techniques

Feature based sequence classification methods:

In feature-based sequence classification methods first sequences are converted into vectors and then these vectors are classified by assigning a suitable class label. Also, note that future based sequence classification methods have many advantages.

In distance-based sequence classification methods a distance metric is defied to find the distance value between any two given data sequences and then distance classifiers such as KNN are used for sequence data classification and sequence data clustering. The sequence data classification task can be defined in a natural way as assigning correctly selected class labels to new data sequences based on the already systematically and conveniently mined knowledge that is gained in the training stage of the classifier. Distance measure play an important role for finding similarity between two given data sequences. Distance based methods generally define a special formula to find actual distance measure between a pair of data sequences. The choice of distance measure that is used for sequence data classification is critical in finding the performances of classifier models.

In model-based sequence classification,probability distributions are assumed on the given sequence data records and sequences belonging to the different classes are generated randomly by using assumed statistical probability distributions and the selected model parameters that must be determined from the training data of sequences of input data records which are normally variable length records. Statistical models and other models such as HMM are generally model based classifiers that are used for sequence data classification.

Some examples for sequences of data and their details are:

TABLE-1 Example of Sequence Data for Classification and Clustering

S.No	Name of the sequence classification
1	Protein sequence clustering
2	Gene sequence
3	Health care monitoring
4	Intrusion detection
5	Protein classification
6	DNA classification
7	Speech recognition
8	Text classification
9	Bioinformatics
10	Speech synthesis
11	Player tracking
12	Machine translation
13	Sentiment analysis
14	Biological sequence classification
15	Document classification
16	Part-of-speech tagging
17	Semantically aware sentiment analysis
18	Classification of query log sequences

Sequence classification is the process of finding a class label for a given sequence. Neural-networks are state-of-the-art sequence classifiers. Automata based sequence classifiers are interpretable. In bio-informatics clustering of biological sequences is really a challenging task.Discrete optimization techniques are used to learn interpretable classifiers. Bayesian inference-based classification technique can be used to classify a sequence of observations. A Decision tree classifier model can also be used as a potential sequence data classifier. Probabilistic models have been extensively used for sequence data classification. Hidden Markov models, neural networks models are also used for sequence data classifications.

Various methods developed for sequence prediction are:

1) Structured perceptron
2) Conditional Random Fields
3) Max-Margin Markov Networks and
4) Structural Support Vector Machines

II. RELATED WORK

Existing tree-based classification algorithms are developed only for processing fixed length records but not suitable for processing variable length records such as sequence data processing. To overcome this limitation two-step procedure is used for discrete sequence classification. In the first step given sequences are converted into fixed length records and in the second step existing techniques are used for this sequence classification or clustering. These existing techniquesare costly in terms of computations, missing information and sequence management. To overcome this problem authors, Z. He et al. [1] proposed a new tree-based sequence data classification procedure in such a way that no information is missed, all sub sequences are included correctly and computational time complexity is reduced to the maximum extent. Proposed technique is applied on many real datasets and experimental results so obtained reveal that the proposed technique is superior in producing high accurate results when compared with the existing techniques and advanced techniques.

Garcia et al. [2] analyzed a set of algorithms called model-based sequences for sequence data clustering and then new techniques are proposed for sequence data clustering based on Kullback-Leibler divergence method. Zhou et al. [3] have measured cohesion and support measures of interesting patterns in sequential patterns with the help of rules and they have given detailed steps for building classifiers. One such technique is classification based on association rule mining. Different machine learning techniques are applied for classification and testing of various patterns. Also experimentally proved that the patterns generated by their proposed techniques are well ordered and define sequences for define classification of patterns. Qiang et al. [4] have reviewed many short text topic modeling techniques including multinomial mixture, global word co-occurrences and self-aggregation. Many state-of-the-art techniques are used for evaluating the performances of may standard datasets. These performances are compared against one another and also against long text topic modelling techniques.

In any data analysis finding outliers is an important task and outlier determination is needed in sequence data processing techniques also. Here, main aim is finding outliers from the given set of sequences. Wang et al [5] proposed a new technique for finding outlier sequences based on average probabilistic strength and different pruning techniques. Spatial data mining and spatial co-location pattern mining are useful operations in some real applications but many of the co-location patterns are redundant. To overcome this problem Wang et al. [6] proposed a new technique called semantic distance between a co-location pattern and its super-pattern. Experiments are conducted on synthetic and real datasets. The results have shown that the proposed method is effective and reduces 50% of redundant patterns.

Fundamental goal of machine learning is finding useful knowledge from the large datasets. It is necessary to find different types of useful patterns based on possible measures such as support, confidence, sequence relationships, longest common sub sequences and many other useful parameters. Utility mining plays a major role in modern applications such as medical diagnosis, business, e-commerce, retail, document management, banking, cross-marketing, finance and research. Gan et al. [7] presented state-of-the-art techniques for mining high-utility patterns based on tree structures, apriori-technique, projection, horizontal and vertical fragments. In depth knowledge of advanced high-utility pattern mining techniques are presented in a systematic way. Exarchos et al. [8] proposed a new two-step algorithm for data sequence classification based on sequence pattern mining and optimization. In the first step a classification model is defined with a set of weights. In the second step to achieve optimal classification accuracy, an elegant optimization technique is applied for optimizing the weights. Experimental results have shown that performance of the proposed method is superior than many existing techniques.

Sequence data processing is necessary in some real contexts. Authors [9] have proposed input output Hidden Markov models for sequence data mining based on some statistical properties. These models are useful for solving grammatical problems. Experimental results have shown that these models have very good generalization features. Markov models are convenient models to represent and manage chronological dependencies present in the sequential data. Chen et al. [10] proposed a new dynamic order Markov model for finding a similarity value between sequences by using a special data structure called probability suffix tree that can find sparse and consecutive patterns. Authors also developed clustering algorithm for clustering categorical sequential data and experimentally verified on real world datasets. Categorical sequential data applications are growing rapidly in business, scientific, and in many day-to-day applications, biological sequences, web clicking, purchasing sequences of objects and so on. Discovering, analyzing, processing and then understanding sequential dependencies is necessary in some applications. Existing sequence similarity finding measures are divided into two types with alignment and without alignment.

Kim and Yue [11] proposed a decision tree framework for learning an accurate spatiotemporal sequence with some attractive features such as scalability, high accuracy, ease of training, ability to handle missing values, test performance is very fast, robust with respect to corrupted data and facility for easy quality checking. Several real datasets are employed for evaluating the performance of the proposed model and performances are compared with the existing decision tree-based sequence learning frameworks. Prediction of contextual sequences is an important task in many applications such as part-of-speech and record identification. One limitation of decision trees is that they cannot represent variable length prediction problems. So, variable length sequences must be converted into simple fixed sequences so that normal decision trees can handle those sequences easily. Decision trees are popularly used for discriminative classification and regression.

B. Sam and R. Huzefa [12] Sequence data classification is an important task in machine learning. Sometimes distance metrics cause information loss in finding distance between a pair of sequences. Authors proposed hidden markov model variant for finding distances between fixed length vectors. Three different algorithms are proposed to find fixed length vectors. These fixed length vectors are useful for classifying sequences of amino acids into structural classes. Different techniques are required for transforming given sequences into another format in such a way that this new format is very convenient for original sequence classification. So, a direct mapping of a given sequence into vectors that can be processed very easily is inevitable. After training with HMM only fixed length vectors will be generated. Authors experimentally evaluated the performance of data for classifying sequences of amino acids into structural classes.

He et al. [13] proposed a reference-based sequence classification frame-work which is treated as generalized pattern-based sequence classification method. In this method reference points and similarity functions are used to find similarity between sequences. Subsets of training sequences are used in creating the reference points and Jaccard similarity function formula is used for similarity computations. Authors have presented reference-based framework consisting of three steps for sequence data classification. The first step extracts alphabets, second step generates candidate sequences and third step selects potential sequences. Proposed method is practically feasible and will be useful to apply in many real time applications. Experiments are conducted on many real time datasets and the results are very good in obtaining high sequence classification accuracies. In the literature, many algorithms have been proposed for clustering discrete sequences.

Shavo et al. [14] proposed automata-based sequence classifiers and their performance is evaluated over the standard bench mark datasets and the experimental results have shown that proposed methods are superior in producing sequence classification accuracies. These proposed methods are highly interpretable over the neural networks particularly with the LSTM-based classifiers.For efficient and effective management of sequential data a measure for comparing sequences is required. Rieck and Laskov [15] proposed a framework of similarity measures for sequences. Nonoverlapping sequential data pattern (NSDP) mining is one of the important types of sequential pattern mining (SPM). NSDPs are able to give only interesting patterns which are effective with reduced search time facilities. The problem with the existing methods is that they produce redundant patterns. To overcome this problem Wu et al. [16] proposed nonoverlapping three-way sequential pattern mining and it reduces time space complexity based on candidate pattern generation technique. For sequence data clustering performance of the proposed method is very high when compared with the existing techniques.

Nonoverlapping sequential pattern mining tasks are important in certain domains. Nonoverlapping Maximal Sequential Patterns (NMSPs) mining is also important and NMSPs are frequent patterns but their super patterns are infrequent. Y. Li et al. [17] proposed a new frequent pattern mining algorithm containing three steps-computing the support with back tracking technique, creating candidate patterns by using join strategy, and then finding non-overlapping maximal sequential patterns. Experiments are conducted on biological sequential datasets. W.Chen et al. [18] have developed a framework for clustering and classification of data sequences. It uses hybrid technique, Hidden Markov models, for clustering and dynamic programming technique for classification. These methods are particularly useful for biological sequence data clustering, classification, and segmentation. Biological sequential structures and patterns are very useful for drug design and disease treatment.

Bioinformatics is an emerging field where sequence data clustering and classification are frequently used. DNA and protein sequences are useful in medical analysis but known details of these sequences are limited. Aleb and Labidi [19] studied DNA sequence data clustering in detail by applying a new method based on standard and very popular K-means clustering algorithm. DNA sequences are used for either clustering or comparative study purpose.Application of deep learning techniques are increasing gradually for sequence data classification with high accuracy. Bao et al. [20] proposed a new binning algorithm for sequence data classification. This algorithm consists of two steps. In the first step, sequences are divided into groups and in the second step groups are transformed into clusters.H. Zheng et al. [21] proposed a sequential data clustering algorithm based on a statistical measure called log-likelihood ratio as distance measure between two sequences. This proposed algorithm is tested on biological sequential datasets and then clustering results are compared with other results of standard clustering algorithms.

Analysis of protein sequences is very useful in certain applications. Yang and Wang [22] have applied techniques for automatic clustering of protein sequences. Clustering of protein sequences is difficult because lack of similarity measures between protein sequences. A new technique was proposed for protein sequence clustering based on statistical properties combined with imprecise probability and suffix tree concepts. Genetic sequence data classification, clustering, and testing are useful in many medical applications. Finding the correct sequence or sub-sequence that is responsible for the identification of disease is also important. Many machine learning techniques are available for medical data analysis particularly decision tree model is a potential and the right model to use for medical data analysis, classification and clustering. Machado et al. [23] proposed a decision tree model for classification of cancer data. Cancer datasets are employed in experimentation and the classification accuracies are determined and compared with other methods. After experimentation they have clearly observed and identified that decision tree model is highly predictive with high classification accuracies and moreover decision tree model is easy to apply for many medical operations.

Siddiquee and Tasnim [24] have analyzed and described thoroughly the standard decision tree (ID3) implementation for classifying DNA sequences. Training and evaluation are carried out on ID3 and Chi-square testing is used for stopping the node splits. Test accuracy is determined by validating on a test dataset. Set of DNA sequences is used as a training dataset and a separate dataset is used for testing the ID3 decision tree classifier model. Statistical techniques are very useful for handling data operations on sequences. In this paper Dietterich [25] has clearly explained machine learning methods such as sliding window methods, HMM, sliding window models, maximum entropy models, and conditional random fields for efficient management of sequences. Also, author has discussed many open issues about classification of sequences. Sequential data learning problems are commonly appearing in many domains.

Sequence data classifications such as finance, retail, information retrieval, document analysis, genomic analysis, protein sequence classification, and health informatics are inevitable in many application domains. Z. Xing et al. [26] presented a detailed survey details of sequence data classifications and they have provided five sequence data types. They revealed that most of the data sequences are based on symbolic data sequences. They said that the problem of complex sequence data classification is still an open problem. Streaming data operations are also really facing many problems.The main aim of nonoverlapping sequential data pattern mining is to find repetitive sequential data patterns. Here, nonoverlapping patterns means that characters present in the same position must be different and overlapping patterns means characters at the same position may be repeatedsometies. Wu et al. [27] proposed a new algorithm called high average utility nonoverlapping sequential data patterns algorithm. It consists of two important steps. The first step is for calculating the support by using depth-first search and backtracking strategies with more advanced nettree data structure and the second step is for reducing the candidate patterns that are already generated by using pattern join strategies. The proposed pattern finding algorithm is efficient and high accurate for pattern data classification and clustering operations and the selected data structure is highly efficient and effective for sequence data operations.

III. PROBLEM DEFINITION

In general sequence data records are variable length records. Main problem in sequence data processing is converting variable length record sequences into fixed length sequence of records. Sequence data classification and clustering are necessary operations in many real-world applications but sequence data consists of variable length sequences of records and processing of these variable length records is very difficult. Therefore, special middle level efficient sequence data conversion processing techniques are required for transforming variable length sequences into fixed length vectors before actually using of many data structures and applying existing fixed length record processing techniques such as decision trees, random forests, ensemble techniques for sequence data classification and clustering.

IV. PROPOSED SEQUENCE DATA SIMILARITY FINDING METHOD

In this paper a new technique is proposed for sequence data similarity (SDS) finding and then this technique is applied for sequence data classification and clustering. Sequence data similarity is computed based on total number of common longest subsequences presented in the two given data sequences, S_1 and S_2. Eachsubsequence is taken as the longest common subsequence. That is, this newly proposedsequence data similarity finding technique is based on the total number of common longest subsequences present between two given data sequences, S_1 and S_2.For a given pair of data sequencesS_1 and S_2 the operation may be either classification or clustering. Both these operations can be executed successfully with the sequence data similarity measure value computed with the similarity finding formula. SDS is applicable between two comparable data sequences but not with un-comparable data sequences.

Sequence data similarity (SDS) between two data sequences (pair o sequences) is defined as

$$SDS(S_1, S_2) = \frac{|LCS1(S_1, S_2)| + |LCS2(S_1, S_2)| + \cdots + |LCSn(S_1, S_2)|}{|S_1| + |S_2| - |LCS1(S_1, S_2)| - |LCS2(S_1, S_2)| - \cdots - |LCSn(S_1, S_2)|} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$LCS1(S_1, S_2) is\ first\ common\ subsequence\ between\ S_1\ and\ S_2$$

$LCS2(S_1, S_2)$ *is second common subsequence between* S_1 *and* S_2

……

……

$LCSn(S_1, S_2)$ *is* n^{th} *common subsequence between* S_1 *and* S_2

In simple terms SDS is defined as $SDS(S_1, S_2) = \frac{z}{x+y-z}$

Where x = length of given first input sequence, y = length of given second input sequence and z = sum of lengths of all common subsequences in both the given input pair of data sequences.

4.1 Sequence Data Similarity Finding EXAMPLE-1

Consider two sample and simple data sequences S_1 = <A,B,C,D> and S_2 = <A, B, C>. Now, the requirement is to find quantitatively the sequence data similarity measure value between two given data sequences, S_1 and S_2 using newly proposed sequence similarity finding technique.

$$SDS(S_1, S_2) = \frac{|LCS1(S_1, S_2)|}{|S_1| + |S_2| - |LCS1(S_1, S_2)|} = \frac{3}{4+3-3} = \frac{3}{4} = 0.75$$

Where $|LCS1(S_1, S_2)|$ is the only one longest subsequence present between two given data sequences S_1 and S_2 and the total number of longest common subsequences = 1. SDS value is 0.75 > 0.5 and 0.5 is the selected similarity threshold value. So, both the input sequences are grouped into the same cluster in the case of clustering operation and in the case of classification, if the class label of $S_1$is 5 then the class label of $S_2$is also assigned as 5.

4.2 Sequence Data Similarity Finding EXAMPLE-2

Assume that S_1 = <A, B, C, D, E, F> and S_2 = <A, B, E, F> are two given data sequences. This pair of input data sequences have two longest common subsequences. That is, here, two distinct longest common subsequences are present in the given two data sequences S_1 and S_2. LCS1(S_1, S_2) =<A, B> is the first longest common subsequence and LCS2(S_1, S_2) =<E, F> is the seconddistinctlongest common subsequence.

$$SDS(S_1, S_2) = \frac{|LCS1(S_1, S_2)| + |LCS2(S_1, S_2)|}{|S_1| + |S_2| - |LCS1(S_1, S_2)| - |LCS2(S_1, S_2)|} = \frac{2+2}{6+4-2-2} = \frac{4}{6} = 0.667$$

In the case of clustering data sequences S_1 and S_2 are grouped into the same cluster because SDS is 0.667 > 0.5. Here, 0.5 is the specified threshold value.In the case of classification both sequences S_1 and S_2 are assigned the same class label. If the class label of $S_1$is 2 then the class label of $S_2$is also assigned 2. So, SDS is used for both classification and clustering.

4.3 Sequence Data Similarity Finding EXAMPLE-3

Assume that S_1 = <A, B, C, D, E, F, G> and S_2 = <A, B, C, E, F, G> are two given data sequences.
Here, two distinct longest common subsequences are present in the given two data sequences S_1 and S_2. <A, B, C> is the first common longest subsequence and <E, F, G> is the second distinct longest common subsequence.

$$SDS(S_1, S_2) = \frac{|LCS1(S_1, S_2)| + |LCS2(S_1, S_2)|}{|S_1| + |S_2| - |LCS1(S_1, S_2)| - |LCS2(S_1, S_2)|} = \frac{3+3}{7+6-3-3} = \frac{6}{7} = 0.857$$

Threshold similarity value 0.5 and SDS is 0.857. So, given pair of input sequences S_1 and S_2 are clustered into the same cluster group and in the case of classification class label of S_1 is assigned as the class label of S_2.

4.4 Sequence Data Similarity Finding EXAMPLE-4

Assume that S_1 = <A, B, C, D, E, F, G, H, I> and S_2 = <A, B, E, F, H, I> are two given data sequences, pair of input data sequences. Here, three distinct longetcommon subsequences are present in the given two data sequences S_1 and S_2. <A, B> is the first longest common subsequence, <E, F> is the second distinct longest subsequence and <H, I> is the third common longest subsequence.

$$SDS(S_1, S_2) = \frac{|LCS1(S_1, S_2)| + |LCS2(S_1, S_2)| + |LCS3(S_1, S_2)|}{|S_1| + |S_2| - |LCS1(S_1, S_2)| - |LCS2(S_1, S_2)| - |LCS3(S_1, S_2)|} = \frac{2+2+2}{9+6-2-2-2} = \frac{6}{9} = 0.667$$

Threshold similarity value 0.5 and SDS is 0.667. So, given pair of input data sequences S_1 and S_2 are clustered into the same cluster group and in the case of sequence data classification class label of S_1 is assigned as the same class label of S_2.

Arithmetic mean (AM), geometric mean (GM) and harmonic mean (HM) can also be applied for sequence data similarity finding validation and verification. In the EXAMPLE-4, let x = SDS(S_1, S_2) be the sequence data similarity and y = 1/total number of distinct common longest subsequences. Here, x and y are two distinct variable parameters.

$$AM(S_1, S_2) = \frac{x+y}{2} = \frac{SDS(S_1, S_2) + y}{2} = \frac{0.667 + \frac{1}{3}}{2} = \frac{0.667 + 0.333}{2} = \frac{0.999}{2} = 0.499$$

$$GM(S_1, S_2) = \sqrt{xy} = \sqrt{0.67 * 0.33} = \sqrt{0.2211} = 0.47$$

$$HM(S_1, S_2) = \frac{2 * x * y}{x+y} = \frac{2 * SDS(S_1, S_2) * y}{SDS(S_1, S_2) + y} = \frac{2 * 0.667 * 0.33}{0.667 + 0.33} = \frac{0.44}{0.997} = \frac{0.999}{2} = 0.441$$

Also, GM^2 = AM * HM value is verified. That is, 0.47 * 0.47 = 0.499 * 0.441. That is, 0.2209 = 0.220

4.5 Sequence Data Similarity Finding EXAMPLE-5

In this example, generally consider once again two sample and very simple given input data sequences S_1 = <A, B, C, D> and S_2 = <A, B, C, D>. Now, the requirement is to find quantitatively the sequence data similarity measure value between two given data sequences, S_1 and S_2 using newly proposed sequence similarity finding technique. remember that both the given input sequences are 100% similar

$$SDS(S_1, S_2) = \frac{|LCS1(S_1, S_2)|}{|S_1| + |S_2| - |LCS1(S_1, S_2)|} = \frac{4}{4+4-4} = \frac{4}{4} = 1.0$$

Where $|LCS1(S_1, S_2)|$ is the only one common longest subsequence presented between two given data sequences S_1 and S_2 and the total number of common longest subsequences = 1. Observe that here sequence data similarity is 1.0 and 1.0 similarity value indicate that given input data sequences are 100% similar. That is similarity value 1.0 indicates 100% similarity. Similarly note that if two given input data sequences have no common longest subsequences or zero common longestsubsequences then their similarity measure is zero (0). For clarity purpose always sequence data similarity values are normalized and correct similarity measure range lies between 0 and 1 both inclusive.

Only threshold satisfied sequence data similarity value measures are taken into consideration and others are discarded. Either sequence data classification or clustering is performed only with similarity measure values higher than the specified threshold values (normally threshold similarity measure value must be greater than 0.5). Longest common subsequence (LCS) maintains or preserves and manages order as well as commonality of longest subsequences. There may exist 'n' number of common longest subsequences in the given two sequence data inputs, pair of input data sequences. All these common longest subsequences and their lengths are used in sequence data similarity computations.

Similarity measure is 1.0, so, pair of given sequences are placed in the same cluster. in the case of classification class label of S_1 is assigned as the class label of S_2. When SDS is 1.0 it is called perfect similarity measure and when SDS is zero given pair of sequences are perfectly dissimilar and not possible to cluster or classify.

CONCLUSIONS

Sequence data operations are plentiful in the society in both normal and stream sequence data operations. In the present paper a new technique is proposed for sequence data similarity finding. This technique is very simple to compute with very low time complexity and very convenient for sequence data operations such as sequence data classification and sequence data clustering. Sequence data similarity is computed based on distinct common subsequences in the given input data sequences. In the future efficient sequence data similarity finding methods will be systematically investigated for betterment of sequence data operations and special intended techniques will be applied to explore and develop a more general framework for effective sequence data classification, clustering and sequence data pattern mining. In the future high performance sequence data similarity finding formulas are strongly and systematically operations will be carried out in order to produce efficient and effective similarity finding formulas.

REFERENCES

[1] Z. He, Z. Wu, G. Xu, Y. Liu an Q. Zou, "Decision Trees for Sequences", Published in: IEEE Transactions on Knowledge and Data Engineering (Volume: 35, Issue: 1, 01 January 2023).

[2] D. Garcia, E. Parrado and E. Diaz, "A New Distance Measure for Model-based Sequence Clustering", Published 1-july-2009, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[3]C. Zhou, B.Cule and B. Goethals, "Pattern Based Sequence Classification", IEEE Transactions on Knowledge and Data Engineering 2015.

[4] Qiang J, Qian Z, Li Y, Yuan Y, Wu X,"Short text topic modelling techniques, Applications, and Performance: A Survey". IEEE Transactions on Knowledge and Data Engineering (TKDE), 2020.

[5] Wang T, Duan L, Dong G, Bao Z,"Efficient mining of outlying sequence patterns for analysingoutlierness of sequence data". ACM Trans KnowlDiscov Data 14(5):62, 2020.

[6] Wang L, Bao X, Zhou L (2018), "Redundancy reduction for prevalent co-location patterns". IEEE Trans Knowl Data Eng 30(1):142–155.

[7] Gan W, Lin JC-W, Zhang J, Fournier-Viger P, Chao H-C, Tseng VS, Yu PS, "A survey of utility-oriented pattern mining". IEEE Transactions on Knowledge and Data Engineering (Volume: 33, Issue: 4, 01 April 2021)

[8] T. P. Exarchos, M. G. Tsipouras, C. Papaloukas, and D. I. Fotiadis, "A two stage methodology for sequence classification based on sequential pattern mining and optimization," Data & Knowledge Engineering, vol. 66, no. 3, pp. 467–487, 2008.

[9] Y. Bengio and P. Frasconi, "Input-output HMM's for sequence processing". Published in: IEEE Transactions on Neural Networks (Volume: 7, Issue: 5, September 1996), 7(5):1231–1249

[10] R. Chen, H. Sun, L. Chen, J. Zhang, and S. Wang, "Dynamic order Markov model for categorical sequence clustering", Journal of Big Data (2021) 8:154 https://doi.org/10.1186/s40537-021-00547-2

[11] T. Kim and Y. Yue, "A Decision Tree Framework for Spatiotemporal Sequence Prediction", KDD'15, August 10-13, 2015, Sydney, NSW, Australia. Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM 978-1-4503-3664-2/15/08 ...$15.00. DOI: http://dx.doi.org/10.1145/2783258.2783356

[12] B. Sam and R. Huzefa, "A Hidden Markov Model Variant for Sequence Classification", Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence.

[13] Z. He, G. Xu, C. Sheng, BO Xu, and Q. Zou, "Reference based sequence classification", IEEE Access, 14-DEC-2020.

[14] M. Shvo, A.C. Li, R. Toro, "Interpretable Sequence Classification via Discrete Optimization", 6-Oct-2020, Department of Computer Science, University of Toronto, Canada.

[15] K. Rieck and P. Laskov, "Linear-time computation of similarity measures for sequential data," Journal of Machine Learning Research, vol. 9, no. 1, pp. 23–48, 2008.

[16] Wu Y, Luo L, Li Y, Guo L, Fournier-Viger P, Zhu X, Wu X (2021) NTP-Miner: Nonoverlapping three-way sequential pattern mining. ACM Transactions on Knowledge Discovery from Data.

[17] Y. Li, S. Zhang, L. Guo, J. Liu, Y. Wu and X. Wu, "Non overlapping maximal sequential pattern mining", Accepted: 7 October 2021 / Published online: 10 January 2022, Applied Intelligence (2022) 52:9861–9884, https://doi.org/10.1007/s10489-021-02912-3

[18] W. Chen, C. Zhang, and X. Chen,"Biological Sequence Clustering and Classification with a Hybrid Method and Dynamic Programming", Published in:21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07), Date Added to IEEE Xplore: 27 Aug 2007

[19] N. Aleb and N. Labidi, "An Improved K-Means Algorithm for DNA Sequence Clustering", Published in: 2015 26th International Workshop on Database and Expert Systems Applications (DEXA), Date Added to IEEE Xplore: 15 February 2016

[20] Bao, Vinh and Hoai, "A Deep Embedded Clustering Algorithm for the Binning of Metagenomic Sequences", Published in: IEEE Access (Volume: 10), Date of Publication: 23 May 2022

[21] H. Zheng, H. Wang and J. Hu, "Cluster Analysis of Regulatory Sequences with a Log Likelihood Ratio Statistics-based Similarity Measure", Published in: 2007 IEEE 7th International Symposium on BioInformatics and BioEngineering, Date Added to IEEE Xplore: 05 November 2007

[22] J. Yang and W. Wang, "Towards automatic clustering of protein sequences", Published in: Proceedings. IEEE Computer Society Bioinformatics Conference, Date Added to IEEE Xplore: 10 December 2002

[23] P. Machado, Gomes, Maia, Stransky and Estefano, "A decision tree to improve identification of pathogenic mutations in clinical practice", Nascimento et al. BMC Medical Informatics and Decision Making (2020) 20:52 https://doi.org/10.1186/s12911-020-1060-0

[24] M. A. Siddiquee and H. Tasnim, "A Comprehensive Study of Decision Trees to Classify DNA Sequences", Article 1 (September 2018), 4 pages. https://doi.org/unassigned AUTHOR CONTRIBUTION

[25] Machine Learning for Sequential Data: A Review Thomas G. Dietterich Oregon State University, Corvallis, Oregon, USA.

[26] Z. Xing, J. Pei, and E. Keogh, "A Brief Survey on Sequence Classification", SIGKDD Explorations Volume 12, Issue 1.

[27] Wu Y, Geng M, Li Y, Guo L, Li Z, Fournier-Viger P, Zhu X, Wu X,"HANP-Miner: High average utility nonoverlapping sequential pattern mining". Knowl-Based Syst 229(107361), 2021.