ABSTRACT

Research on sequence data concentrates on the discovery of frequently occurring patterns. However, comparatively less amount of work has been carried out in the area of sequence data classification and clustering.

The contribution of this thesis is in the development of new methods for classification and clustering of sequential data. The thesis introduces solutions for real-world applications dealing with classification of system call sequences in intrusion detection and clustering of web navigational sequences in web usage mining. We have chosen DARPA'98 IDS benchmark dataset for the sequence classification experiments. For sequence clustering experiments, msnbc web navigational benchmark dataset was chosen.

Initially, we established the hypothesis that while comparing two sequences, considering only the order or the content information embedded in the two sequences results in poor performance in classification and clustering tasks. kNN classification and PAM clustering algorithms are utilized for classification and clustering tasks, respectively and sliding window technique is used for extracting subsequences. To establish the hypothesis four distance/similarity measures namely, Euclidean, Cosine, Jaccard and Binary Weighted Cosine measures were used along with various sliding window sizes.

Based on the results from our hypothesis testing, we designed a new similarity metric, $S^3M$ which considers both the order of occurrence as well as the content information of the two sequences being compared. Better classification accuracy and clustering quality were achieved with the newly devised measure.

In clustering of sequences, the goodness of the clusters was measured using average levenshtein distance. We propose a new partitional clustering algorithm for sequence data, SeqPAM. SeqPAM differs from PAM in medoid selection as well as the optimization function. The superiority of the proposed algorithm was demonstrated over PAM clustering algorithm.

Further, a new indiscernibility-based rough agglomerative hierarchical clustering algorithm for sequential data is proposed. Here, the indiscernibility relation has been extended to a tolerance relation with the transitivity property being relaxed. Initial clusters are formed using a similarity upper approximation. Subsequent clusters are formed using the concept of constrained-similarity upper approximation wherein a condition of relative similarity is used as a merging criterion. We compared the results of the proposed approach with that of the traditional hierarchical clustering algorithm using vector coding of sequences.