**A heuristic approach for finding similarity indexes of multivariate data sets**

ABSTRACT

Multivariate data sets (MDSs), with enormous size and certain ratio of noise/outliers, are generated routinely in various application domains. A major issue, tightly coupled with these MDSs, is how to compute their similarity indexes with available resources in presence of noise/outliers - which is addressed with the development of both classical and non-metric based approaches. However, classical techniques are sensitive to outliers and most of the non-classical approaches are either problem/application specific or overlay complex. Therefore, the development of an efficient and reliable algorithm for MDSs, with minimum time and space complexity, is highly encouraged by the research community. In this paper, a non-metric based similarity measure algorithm, for MDSs, is presented that solves the aforementioned issues, particularly, noise and computational time, successfully. This technique finds the similarity indexes of noisy MDSs, of both equal and variable sizes, through utilizing minimum possible resources i.e., space and time. Experiments were conducted with both benchmark and real time MDSs for evaluating the proposed algorithm`s performance against its rival algorithms, which are traditional dynamic programming based and sequential similarity measure algorithms. Experimental results show that the proposed scheme performs exceptionally well, in terms of time and space, than its counterpart algorithms and effectively tolerates a considerable portion of noisy data.

**Keyword:** Similarity index; Multivariate data set; Outliers; The longest common subsequence