World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:14, No:12, 2020

Distances over Incomplete Diabetes and Breast Cancer Data Based on Bhattacharyya Distance

Authors : Loai AbdAllah, Mahmoud Kaiyal

Abstract : Missing values in real-world datasets are a common problem. Many algorithms were developed to deal with this problem, most of them replace the missing values with a fixed value that was computed based on the observed values. In our work, we used a distance function based on Bhattacharyya distance to measure the distance between objects with missing values. Bhattacharyya distance, which measures the similarity of two probability distributions. The proposed distance distinguishes between known and unknown values. Where the distance between two known values is the Mahalanobis distance. When, on the other hand, one of them is missing the distance is computed based on the distribution of the known values, for the coordinate that contains the missing value. This method was integrated with Wikaya, a digital health company developing a platform that helps to improve prevention of chronic diseases such as diabetes and cancer. In order for Wikaya’s recommendation system to work distance between users need to be measured. Since there are missing values in the collected data, there is a need to develop a distance function distances between incomplete users profiles. To evaluate the accuracy of the proposed distance function in reflecting the actual similarity between different objects, when some of them contain missing values, we integrated it within the framework of k nearest neighbors (kNN) classifier, since its computation is based only on the similarity between objects. To validate this, we ran the algorithm over diabetes and breast cancer datasets, standard benchmark datasets from the UCI repository. Our experiments show that kNN classifier using our proposed distance function outperforms the kNN using other existing methods.

Keywords : missing values, incomplete data, distance, incomplete diabetes data
Conference Title : ICSRD 2020 : International Conference on Scientific Research and Development
Conference Location : Chicago, United States
Conference Dates : December 12-13, 2020