ABSTRACT

Various methods of machine learning have been implemented in the medical field to classify various diseases, such as diabetes. The *k*-nearest neighbors (KNN) is one of the most known approaches for predicting diabetes. Many researchers have found by combining KNN with one or more other algorithms may provide a better result. In this paper, a combination of three procedures, removing noise, reducing the dimension, and weighting distance, is proposed to improve a standard voting-based KNN to classify Pima Indians Diabetes Dataset (PIDD) into two classes. First, the noises in the training set are removed using *k*-means clustering (KMC) to make the voter data in both classes more competent. Second, its dimensional is then reduced to decrease the intra-class data distances but increase the inter-class ones. Two methods of dimensional reduction: principal component analysis (PCA) and autoencoder (AE), are applied to investigate the linearity of the dataset. Since there is an imbalance on the dataset, a proportional weight is incorporated into the distance formula to get the fairness of the voting. A 5-fold cross validation-based evaluation shows that each proposed procedure works very well in enhancing the KNN. KMC is capable of increasing the accuracy of KNN from 81.6% to 86.7%. Combining KMC and PCA improves the KNN accuracy to be 90.9%. Next, a combination of KMC and AE enhances the KNN to gives an accuracy of 97.8%. Combining three proposed procedures of KMC, PCA, and Weighted KNN (WKNN) increases the accuracy to be 94.5%. Finally, the combination of KMC, AE, and WKNN reaches the highest accuracy of 98.3%. The facts that AE produces higher accuracies than PCA inform that the features in the dataset have a high non-linearity.

Keywords—autoencoder, diabetes, k-means clustering, k-nearest neighbors, principal component analysis