Implementasi K-harmonic means clustering untuk imputasi missing values = Implementation of K-harmonic means clustering for missing values imputation

Taufik Anwar, author

Abstrak

Pembersihan data merupakan salah satu langkah dalam preprocessing yang dalam prosesnya sering menemukan nilai yang hilang dalam dataset. Nilai yang hilang adalah suatu kondisi di mana tidak ada nilai untuk pengamatan. Langkah cepat yang dapat diambil untuk menangani nilai yang hilang adalah menghapus pengamatan yang mengandung nilai yang hilang, tetapi ini dapat mengurangi informasi dalam data. Cara lain untuk menangani nilai yang hilang adalah dengan menggunakan imputasi dengan mean, median, atau mode nilai dalam variabel di mana nilai-nilai yang hilang berada, dan beberapa metode imputasi seperti imputasi dengan pendekatan clustering. Imputasi dengan pendekatan clustering adalah fokus dari penelitian ini, di mana penelitian ini menggunakan K-Harmonic Means yang telah disesuaikan untuk menangani data numerik dan kategorik campuran. K-Harmonic Means adalah perpanjangan dari K-Means dengan mengurangi masalah sensitivitas inisialisasi centroid acak. Imputasi nilai-nilai yang hilang dilakukan dengan mendistribusikan pengamatan yang memiliki nilai-nilai yang hilang ke cluster dan mengganti nilai-nilai yang hilang dengan informasi centroid pada cluster yang sama. Simulasi menggunakan data dengan nilai-nilai yang hilang yang dibuat menggunakan mekanisme yang hilang sepenuhnya secara acak dengan proporsi 10%, 15%, dan 20% dari total pengamatan. Hasil simulasi dievaluasi menggunakan root mean square error (RMSE) dan nilai akurasi masing-masing nilai imputasi untuk data numerik dan kategorikal. Dalam penelitian ini, hasil imputasi optimal diperoleh pada data dengan proporsi nilai yang hilang 10%, yang memiliki nilai RMSE rendah dan nilai akurasi tinggi.
<hr>
Data cleaning is one step in preprocessing which in the process often finds missing values in the dataset. Missing value is a condition where there is no value for observation. A quick step that can be taken to handle missing values is to delete observations that contain missing values, but this can reduce the information in the data. Another way to handle missing values is to use imputations with the mean, median, or value modes in the variable where the missing values are located, and some imputation methods such as imputation with the clustering approach. Imputation with the clustering approach is the focus of this study, where this study uses K-Harmonic Means that have been adjusted to handle numerical and mixed categorical data. K-Harmonic Means is an extension of K-Means by reducing the sensitivity problem of random centroid initialization. The imputation of missing values is carried out by distributing observations that have missing values to the cluster and replacing the missing values with centroid information on the same cluster. The simulation uses data with missing values that are made using a completely random missing mechanism with a proportion of 10%, 15%, and 20% of the total observations. Simulation results are evaluated using the root mean square error (RMSE) and the accuracy value of each imputation value for numerical and categorical data. In this study, the optimal imputation results are obtained on data with a proportion of missing values of

10%, which has a low RMSE value and a high accuracy value.