Clearing and preparation of data by using clustering methods

nastaran karimi ghasem abad*,Azamossadat Nourbakhsh, Azamossadat Nourbakhsh,

Cleansing errors and noise data is one of the main parts of data maintenance. Data cleansing refers to the consequence of the conducted operations with the purpose of improving quality of datasets. In fact, data cleansing is to decrease and remove copies in dataset. A problem occurs in applied applications of the datasets and will be worse when the source data will be combined. Therefore, data cleansing in considered as the main part if data uniformity. Still, numerous problems and challenges such as error correction, maintaining cleansed data and data cleansing exist in uniform virtual environments regarding data cleansing. In order to solve the challenges, an appropriate framework in needed. Therefore, in this study, a preparation and cleansing method based on clustering has been presented the raw data are clustered because help the cleansing and normalization. Then, data cleansing operation and finally data conversion are implemented. In this study, the experiments were considered in the worse noise condition in order to obtain the best results for the proposed method. Generally, the proposed method shows a data cleansing framework to improve the accuracy of data mining methods and the main advantage of this method is that do not require updating in dynamic environments and by placing each new data in appropriate category, updates the source. In the proposed model, the dataset with 20000 letters was used that after implementing the proposed steps on dataset, two sets were cleansed and compared with the proposed model and according to the results, the proposed model improved the KStar classification algorithm by 25%.

Keywords : cleansing, clustering, data, noise, error