Multi-Level Air Quality Classification in China Using Information Gain and Support Vector Machine

Authors : Bingchun Liu, Pei-Chann Chang, Natasha Huang, Dun Li

Abstract : Machine Learning and Data Mining are the two important tools for extracting useful information and knowledge from large datasets. In machine learning, classification is a wildly used technique to predict qualitative variables and is generally preferred over regression from an operational point of view. Due to the enormous increase in air pollution in various countries especially China, Air Quality Classification has become one of the most important topics in air quality research and modelling. This study aims at introducing a hybrid classification model based on information theory and Support Vector Machine (SVM) using the air quality data of four cities in China namely Beijing, Guangzhou, Shanghai and Tianjin from Jan 1, 2014 to April 30, 2016. China&#39;s Ministry of Environmental Protection has classified the daily air quality into 6 levels namely Serious Pollution, Severe Pollution, Moderate Pollution, Light Pollution, Good and Excellent based on their respective Air Quality Index (AQI) values. Using the information theory, information gain (IG) is calculated and feature selection is done for both categorical features and continuous numeric features. Then SVM Machine Learning algorithm is implemented on the selected features with cross-validation. The final evaluation reveals that the IG and SVM hybrid model performs better than SVM (alone), Artificial Neural Network (ANN) and K-Nearest Neighbours (KNN) models in terms of accuracy as well as complexity.

Keywords : machine learning, air quality classification, air quality index, information gain, support vector machine, cross-validation

Conference Title : ICDMBDDDM 2019 : International Conference on Data Mining, Big Data, Database and Data Management

Conference Location : Sydney, Australia

Conference Dates : January 30-31, 2019