World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:10, No:02, 2016

A Novel Heuristic for Analysis of Large Datasets by Selecting Wrapper-Based Features

Authors : Bushra Zafar, Usman Qamar

Abstract : Large data sample size and dimensions render the effectiveness of conventional data mining methodologies. A data mining technique are important tools for collection of knowledgeable information from variety of databases and provides supervised learning in the form of classification to design models to describe vital data classes while structure of the classifier is based on class attribute. Classification efficiency and accuracy are often influenced to great extent by noisy and undesirable features in real application data sets. The inherent natures of data set greatly masks its quality analysis and leave us with quite few practical approaches to use. To our knowledge first time, we present a new approach for investigation of structure and quality of datasets by providing a targeted analysis of localization of noisy and irrelevant features of data sets. Machine learning is based primarily on feature selection as pre-processing step which offers us to select few features from number of features as a subset by reducing the space according to certain evaluation criterion. The primary objective of this study is to trim down the scope of the given data sample by searching a small set of important features which may results into good classification performance. For this purpose, a heuristic for wrapper-based feature selection using genetic algorithm and for discriminative feature selection an external classifier are used. Selection of feature based on its number of occurrence in the chosen chromosomes. Sample dataset has been used to demonstrate proposed idea effectively. A proposed method has improved average accuracy of different datasets is about 95%. Experimental results illustrate that proposed algorithm increases the accuracy of prediction of different diseases.

Keywords : data mining, generic algorithm, KNN algorithms, wrapper based feature selection
Conference Title : ICDMA 2016 : International Conference on Data Mining and Applications
Conference Location : Melbourne, Australia
Conference Dates : February 04-05, 2016